Envelopes for Efficient Multivariate Parameter Estimation

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xin Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

R. Dennis Cook, Advisor

April, 2014

Acknowledgements

This thesis would not have been possible without the encouragement, help and support of my family, friends and so many wonderful people that I met at the University of Minnesota. My heartfelt thanks to all of you.

I would like to offer my special thanks to Professor Cook for his constant guidance and patience. I feel extremely fortunate to have Dennis as my PhD advisor and feel really lucky to be trained by such a great master in Statistics. Whenever I felt insufficiently intelligent and insufficiently motivated, Dennis helped me to get what I wanted with his vision, insight and enthusiasm. I have benefited greatly from his wisdom, probably much more than I have realized.

I would also like to extend my gratitude to my friends who have come through Professor Cook's group for setting good examples for me: Xin Chen, Liliana Forzani, Do Hyang Kim, Lexin Li and Zhihua Su.

I also owe special thanks to the members of my committee: Professors Christopher Nachtsheim, Adam Rothman and Yuhong Yang for their insightful comments and suggestions over the years.

I would like to thank the faculty and staff in the School of Statistics for their lectures and assistance. In particular, thank Professor Galin Jones for recruiting me as a PhD student; thank Professor Lan Wang for teaching me three core statistics courses, which helped me making my mind to become a statistician; thank Professors Gongjun Xu and Hui Zou for their help with job application. I am also grateful to Professor Wei Pan from Biostatistics Department for serving on my preliminary exam committee.

Again, I thank all my friends. I find words can hardly describe their incredible impact on my life. Finally, I want to thank my parents and my wife Qing Mai. Without their support and love, it would be meaningless and certainly impossible for me to finish this work.

Dedication

To my parents.

Abstract

Multivariate linear regression with predictors $\mathbf{X} \in \mathbb{R}^p$ and responses $\mathbf{Y} \in \mathbb{R}^r$ is a cornerstone of multivariate statistics. When p and r are not small, it is widely recognized that reducing the dimensionalities of $\mathbf{X}$ and $\mathbf{Y}$ may often result in improved performance. Cook, Li and Chiaromonte (2010) proposed a new statistical concept–envelopes for increasing efficiency in estimation and prediction in multivariate linear regression. The idea is to envelope the information in the data that is material to the estimation of the parameters of interest, while excluding the information that is immaterial to estimation. This is achieved by estimating an envelope, which is essentially a targeted dimension reduction subspace for particular parameters of interest, to reduce the dimensionality of original problems. In this dissertation, we first propose a fast and stable 1D algorithm for envelope estimation in general. Because envelope estimation involves Grassmann manifold optimizations, our scalable algorithm largely lessens the computational burdens of past and future envelope methods. We then naturally propose two new envelope methods for simultaneously reducing $\mathbf{X}$ and $\mathbf{Y}$, and for combining envelopes with reduced-rank regression. At the final chapter, we extend the idea of envelope beyond multivariate linear model to rather arbitrary multivariate estimation problems. We propose a constructive definition and a unified framework for incorporating envelopes with many future applications.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and organization of this dissertation

An important topic in statistics is to reduce the dimensionality of data set and parameter space without losing any necessary information. There is a nascent area in dimension reduction called envelopes, whose goal is to increase efficiency in multivariate parameter estimation and prediction. This is achieved by enveloping the information in the data that is material to the estimation of the parameters of interest, while excluding the information that is immaterial to estimation. The reduction in estimative variation can be quite substantial when the immaterial variation is relatively large.

Envelopes were introduced by Cook, Li and Chiaromonte (2010) for response reduction in the multivariate linear model with normal errors. They showed that the asymptotic covariance matrix of the envelope estimator of the regression coefficients is never larger than that of the usual maximum likelihood estimator and has the potential to be substantially smaller. When some predictors are of special interest, Su and Cook (2011) proposed the partial envelope model, with the goal of improving efficiency of the estimated coefficients corresponding to these particular predictors. Cook, Helland and Su (2013) used envelopes to study predictor reduction in multivariate linear regression and established a connection between envelopes and the SIMPLS algorithm (de Jong 1993; see also ter Braak and de Jong , 1998) for partial least squares regression. They showed that SIMPLS is based on a $\sqrt{n}$-consistent estimator of an envelope and, using this connection, they proposed an envelope estimator that has the potential to dominate SIMPLS in prediction. Still in the context of multivariate linear regression, Schott (2013) used saddle point approximations to improve a likelihood ratio test for the envelope dimension. Su and Cook (2013) adapted envelopes for the estimation of multivariate means with heteroscedastic errors, and Su and Cook (2012) introduced a different type of envelope construction, called inner envelopes, that can produce efficiency gains when envelopes offer no gains.

This dissertation deepens and broadens existing statistical theories and methodologies

in the envelope literature by making the following major achievements: (i) a fast and stable one-dimensional algorithm is proposed in Chapter 2 for estimating an envelope in general; (ii) in Chapter 3, we connect envelope methods with other popular methods for dimension reduction, and introduce envelopes for simultaneously reducing the predictors and responses in multivariate linear regression; (iii) in Chapter 4, we construct a hybrid method of reduced-rank regression and envelope models, and also illuminate such general applicable approach of combining the strength of envelopes and other methods; (iv) finally in Chapter 5, we break through the limitation of envelopes in multivariate linear models and introduce the novel constructive principle of enveloping an arbitrary parameter vector or matrix, based on a pre-specified asymptotically normal estimator.

1.2 Notations

The following notations and definitions will be used in our exposition.

Matrices and subspaces.

Let $\mathbb{R}^{m\times n}$ be the set of all real $m \times n$ matrices and let $\mathbb{S}^{k\times k}$ be the set of all real and symmetric $k \times k$ matrices. Suppose $\mathbf{M} \in \mathbb{R}^{m\times n}$, then $\text{span}(\mathbf{M}) \subseteq \mathbb{R}^m$ is the subspace spanned by columns of $\mathbf{M}$. The Grassmann manifold consisting of the set of all u dimensional subspaces of $\mathbb{R}^r$, $u \leq r$, is denoted as $\mathcal{G}_{u,r}$. We use $\mathbf{P}_{\mathbf{A}(\mathbf{V})} = \mathbf{A}(\mathbf{A}^T\mathbf{V}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}$ to denote the projection onto $\text{span}(\mathbf{A})$ with the $\mathbf{V}$ inner product and use $\mathbf{P}_{\mathbf{A}}$ to denote projection onto $\text{span}(\mathbf{A})$ with the identity inner product. Let $\mathbf{Q}_{\mathbf{A}(\mathbf{V})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{V})}$. We will use operators vec : $\mathbb{R}^{a\times b} \to \mathbb{R}^{ab}$, which vectorizes an arbitrary matrix by stacking its columns, and vech : $\mathbb{R}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$, which vectorizes a symmetric matrix by stacking the unique elements of its columns. Let $\mathbf{A} \otimes \mathbf{B}$ denote the Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$. The Kronecker product of two subspaces $\mathcal{A}$ and $\mathcal{B}$ is defined as $\mathcal{A} \otimes \mathcal{B} = \{\mathbf{a} \otimes \mathbf{b} | \mathbf{a} \in \mathcal{A},\ \mathbf{b} \in \mathcal{B}\}$, which equals $\text{span}(\mathbf{A} \otimes \mathbf{B})$ for any $\mathbf{A}$ and $\mathbf{B}$ such that $\mathcal{A} = \text{span}(\mathbf{A})$ and $\mathcal{B} = \text{span}(\mathbf{B})$. For an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$, their direct sum is defined as the $(m+p) \times (n+q)$ block diagonal matrix $\mathbf{A} \oplus \mathbf{B} = \text{diag}(\mathbf{A}, \mathbf{B})$. We will also use the $\oplus$ operator for two subspaces. If $\mathcal{S} \subseteq \mathbb{R}^p$ and $\mathcal{R} \subseteq \mathbb{R}^q$ then $\mathcal{S} \oplus \mathcal{R} = \text{span}(\mathbf{S} \oplus \mathbf{R})$ where $\mathbf{S}$ and $\mathbf{R}$ are basis matrices for $\mathcal{S}$ and $\mathcal{R}$. The sum of two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}^m$ is defined as $\mathcal{S}_1 + \mathcal{S}_2 = \{\mathbf{v}_1 + \mathbf{v}_2 | \mathbf{v}_1 \in \mathcal{S}_1, \mathbf{v}_2 \in \mathcal{S}_2\}$.

Random vectors and their distributions.

For three arbitrary random vectors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, let $\mathbf{A} \sim \mathbf{B}$ denote that $\mathbf{A}$ has the same distribution as $\mathbf{B}$, let $\mathbf{A} \perp\!\!\!\perp \mathbf{B}$ denote that $\mathbf{A}$ is independent of $\mathbf{B}$ and let $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ indicate that $\mathbf{A}$ is conditionally independent of $\mathbf{B}$ given $\mathbf{C}$. In multivariate linear regression of $\mathbf{Y}$ on $\mathbf{X}$: $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$, we use $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}}$ to denote the population covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, use $\boldsymbol{\Sigma}_{\mathbf{XY}}$ to denote their cross-covariance matrix, and use $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ to denote the covariance matrix of the population residual vector $\boldsymbol{\epsilon} \sim \mathbf{Y}|\mathbf{X}$. Then

$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}$. The sample counterparts of these population covariance matrices are denoted by $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{Y}}$, $\mathbf{S}_{\mathbf{XY}}$, $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$. Sample covariance matrices based on an i.i.d. sample of $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ are defined with the divisor n. For instance, $\mathbf{S}_{\mathbf{X}} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T/n$, $\mathbf{S}_{\mathbf{XY}} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^T/n$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ denotes the covariance matrix of the residuals from the linear fit of $\mathbf{Y}$ on $\mathbf{X}$: $\mathbf{S}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{S}_{\mathbf{XY}}$, and $\mathbf{S}_{\mathbf{Y} \circ \mathbf{X}} = \mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{S}_{\mathbf{XY}}$ denotes the sample covariance matrix of the fitted vectors from the linear fit of $\mathbf{Y}$ on $\mathbf{X}$. We also define $\boldsymbol{\Sigma}_{\mathbf{A}|\mathbf{B}}$ and $\mathbf{S}_{\mathbf{A}|\mathbf{B}}$ in the same way for two arbitrary random vectors.

Estimation methods and some common abbreviations.

We use $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}}$ to denote an estimator of $\boldsymbol{\theta}$ with known parameters $\boldsymbol{\alpha}$. If $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a normal random vector with mean 0 and covariance matrix $\boldsymbol{\Phi}$ we write its asymptotic covariance matrix as $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\theta}}) = \boldsymbol{\Phi}$. Some commonly used abbreviations are:

AIC Akaike information criterion

BIC Bayesian information criterion

CCA Canonical correlation analysis

MLE Maximum likelihood estimator

OLS Ordinary least squares

PCA Principal component analysis

PLS Partial least squares. Also, SIMPLS and NIPALS are two popular algorithms for partial least squares (linear) regression, and IRPLS is the popular algorithm for partial least squares in generalized linear models.

RRR Reduced-rank regression

For a common parameter $\boldsymbol{\theta}$ in different models, we will use subscripts to distinguish the estimators according to different models: $\widehat{\boldsymbol{\theta}}_{\text{env}}$ for the envelope estimator, $\widehat{\boldsymbol{\theta}}_{\text{OLS}}$ for the ordinary least square estimator, $\widehat{\boldsymbol{\theta}}_{\text{RE}}$ for the reduced-rank envelope estimator (a method proposed in Chapter 4) and $\widehat{\boldsymbol{\theta}}_{\text{RR}}$ for the reduced-rank regression estimator.

1.3 Envelope models and methods

In this section, we first review some key definition for envelopes and then illustrate the concepts by a simple data set from Kenward's study (1987).

1.3.1 Definition of an envelope

This following definition of a reducing subspace is equivalent to the usual definition found in functional analysis (Conway 1990) and in the literature on invariant subspaces, but the underlying notion of reduction is incompatible with how it is usually understood in statistics. Nevertheless, it is common terminology in those areas and is the basis for the definition of an envelope (Cook, et al., 2010) which is central to our developments.

Definition 1.1. *A subspace $\mathcal{R} \subseteq \mathbb{R}^d$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{d\times d}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_\mathcal{R}\mathbf{M}\mathbf{P}_\mathcal{R} + \mathbf{Q}_\mathcal{R}\mathbf{M}\mathbf{Q}_\mathcal{R}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

The next definition shows how to construct an envelope in terms of reducing subspaces.

Definition 1.2. *Let $\mathbf{M} \in \mathbb{S}^d$ and let $\mathcal{B} \subseteq \text{span}(\mathbf{M})$. Then the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_\mathbf{M}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.*

The intersection of two reducing subspaces of $\mathbf{M}$ is still a reducing subspace of $\mathbf{M}$. This means that $\mathcal{E}_\mathbf{M}(\mathcal{B})$, which is unique by its definition, is the smallest reducing subspace containing $\mathcal{B}$. Also, the $\mathbf{M}$-envelope of $\mathcal{B}$ always exist because of the requirement $\mathcal{B} \subseteq \text{span}(\mathbf{M})$. If $\text{span}(\mathbf{U}) = \mathcal{B}$, then we write $\mathcal{E}_\mathbf{M}(\mathbf{U}) := \mathcal{E}_\mathbf{M}(\text{span}(\mathbf{U})) = \mathcal{E}_\mathbf{M}(\mathcal{B})$ to avoid notation proliferation. Let $\mathcal{E}^\perp_\mathbf{M}(\mathbf{U})$ denote the orthogonal complement of $\mathcal{E}_\mathbf{M}(\mathbf{U})$.

The following proposition summarizes some key algebraic properties of envelopes in Cook et al. (2010). For a matrix $\mathbf{M} \in \mathbb{S}^{p\times p}$, let λ_i and $\mathbf{P}_i$, $i = 1, \ldots, q$, be its distinct eigenvalues and corresponding eigen-projections so that $\mathbf{M} = \sum_{i=1}^q \lambda_i \mathbf{P}_i$. Define the function $f^* : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ as $f^*(\mathbf{M}) = \sum_{i=1}^q f(\lambda_i)\mathbf{P}_i$, where $f : \mathbb{R} \to \mathbb{R}$ is a function such that $f(x) = 0$ if and only if $x = 0$.

Proposition 1.1. *(Cook et al. 2010)*

1. *If $\mathbf{M} \in \mathbb{S}^{p\times p}$ has $q \leq p$ eigenspaces, then the $\mathbf{M}$-envelope of $\mathcal{B} \subseteq \text{span}(\mathbf{M})$ can be constructed as $\mathcal{E}_\mathbf{M}(\mathcal{B}) = \sum_{i=1}^q \mathbf{P}_i\mathcal{B}$;*
2. *With f and f^* as previously defined, $\mathcal{E}_\mathbf{M}(\mathcal{B}) = \mathcal{E}_\mathbf{M}(f^*(\mathbf{M})\mathcal{B})$.*
3. *If f is strictly monotonic then $\mathcal{E}_\mathbf{M}(\mathcal{B}) = \mathcal{E}_{f^*(\mathbf{M})}(\mathcal{B}) = \mathcal{E}_\mathbf{M}(f^*(\mathbf{M})\mathcal{B})$.*

From this proposition, we see that the $\mathbf{M}$-envelope of $\mathcal{B}$ is the sum of the eigenspaces of $\mathbf{M}$ that are not orthogonal to $\mathcal{B}$; that is, the eigenspaces of $\mathbf{M}$ onto which $\mathcal{B}$ projects non-trivially. This implies that the envelope is the span of some subset of the eigenspaces of $\mathbf{M}$. In the regression context, $\mathcal{B}$ is typically the span of a regression coefficient matrix or a matrix of cross-covariances, and $\mathbf{M}$ is chosen as a covariance matrix which is usually positive definite. We next illustrate the potential gain of envelope method using a linear regression example.

1.3.2 Concepts and methodology

We use Kenward's (1987) data to illustrate the working mechanism of envelopes in multivariate linear regression. These data came from an experiment to compare two treatments for the control of an intestinal parasite in cattle. Thirty animals were randomly assigned to each of the two treatments. Their weights (in kilograms) were recorded at the beginning of the study prior to treatment application and at 10 times during the study corresponding to weeks 2, 4, 6, ..., 18 and 19; that is, at two-weeks intervals except the last which was over a one-week interval. The goal was to find if there is a detectable difference between the two treatments and, if such a difference exists, the time at which it first occurred. As emphasized by Kenward (1987), although these data have a typical longitudinal structure, the nature of the disease means that growth during the experiment is not amenable to modeling as a smooth function of time, and that fitting growth profiles with a low degree polynomial may hide interesting features of the data because the mean growth curves for the two treatment groups are very close relative to their variation from animal to animal. Indeed, profile plots of the data suggest no difference between the treatments. Kenward modeled the data using a multivariate linear model with an "ante-dependence" covariance structure. Here we proceed with an envelope analysis based on a multivariate linear model, following the structure outlined by Cook et al. (2010).

Neglecting the basal measurement for simplicity, let $\mathbf{Y}_i \in \mathbb{R}^{10}$, $i = 1, \ldots, 60$, be the vector of weight measurements of each animal over time and let $X_i = 0$ or 1 indicate the two treatments. Our interest lies in the regression coefficient $\boldsymbol{\beta}$ from the multivariate linear regression $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\epsilon}$, where it is assumed that $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_{\mathbf{Y}|X})$. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$, which is also the maximum likelihood estimator. The estimates and their residual bootstrap standard errors are shown in Table 1.1. The maximum absolute t-value over the elements of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is 1.30, suggesting that the treatments do not have a differential affect on animal weight. However, with a value of 26.9 on 10 degrees of freedom, the likelihood ratio statistics for the hypothesis $\boldsymbol{\beta} = 0$ indicates otherwise. We next turn to an envelope analysis.

Let $\boldsymbol{\Gamma} \in \mathbb{R}^{10\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|X}}(\boldsymbol{\beta})$, the $\boldsymbol{\Sigma}_{\mathbf{Y}|X}$-envelope of span($\boldsymbol{\beta}$), and let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be an orthogonal matrix. Then span($\boldsymbol{\beta}$) $\subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|X}}(\boldsymbol{\beta})$ by Definition 1.2 and we can express $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times 1}$ carries the coordinates of $\boldsymbol{\beta}$ relative to the basis $\boldsymbol{\Gamma}$ and $1 \leq u \leq 10$. Also, because span($\boldsymbol{\Gamma}$) is a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{Y}|X}$ (Definition 1.1), the envelope version of the multivariate linear model can now be written as $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta} X + \boldsymbol{\epsilon}$, with $\boldsymbol{\Sigma}_{\mathbf{Y}|X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(10-u)\times(10-u)}$ are positive definite matrices. Under this model, $\boldsymbol{\Gamma}_0^T\mathbf{Y}|X \sim \boldsymbol{\Gamma}_0^T\mathbf{Y}$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}^T\mathbf{Y}|X$. Consequently, $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ does not respond to changes in X either marginally or because of an association with $\boldsymbol{\Gamma}^T\mathbf{Y}$. For these reasons we regard $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ as the immaterial

information and $\boldsymbol{\Gamma}^T\mathbf{Y}$ as the material information. Envelope analyses are particularly effective when the immaterial variation $\text{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y})$ is large relative to the material variation $\text{var}(\boldsymbol{\Gamma}^T\mathbf{Y})$. After finding a value $\widehat{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Gamma}$ that minimizes the likelihood-based Grassmann objective function $\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}}^{-1}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{Y}|X}\boldsymbol{\Gamma}|$, which will be discussed in Section 2.2, over all semi-orthogonal matrices $\boldsymbol{\Gamma} \in \mathbb{R}^{10\times u}$, the envelope estimator of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}}_{\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. Because $u(10-u) \leq 25$, the real dimensions involved in this optimization are small and the `envlp` code can be used without running into computational issues. Standard methods like BIC and likelihood ratio testing can be used to guide the choice of the envelope dimension u. Both methods indicate clearly that $u = 1$ in this illustration. In other words, the treatment difference is manifested in only one linear combination $\boldsymbol{\Gamma}^T\mathbf{Y}$ of the response vector.

The envelope estimate $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is shown in Table 1.1 along with bootstrap standard errors and standard errors obtained from the asymptotic normal distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{env}} - \boldsymbol{\beta})$ by the plug-in method (See Cook et al. (2010) for the asymptotic covariance matrix). We see that the asymptotic standard errors are a bit smaller than the bootstrap standard errors. Using either set of standard errors and using a Bonferroni adjustment for multiple testing, we see that there is a difference between the treatments and that the difference is first manifested around week 10 and remains thereafter. As shown in the final row of Table 1.1, the bootstrap standard errors for the elements of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ were 2.2 to 5.9 times those of $\widehat{\boldsymbol{\beta}}_{\text{env}}$. Hundreds of additional samples would be needed to reduce the standard errors of the elements of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ by these amounts.

We conclude this example by considering the regression of the 6th and 7th element of $\mathbf{Y}$, corresponding to weeks 12 and 14, on X, now letting $\mathbf{Y} = (Y_6, Y_7)^T$. This allows us to represent the regression graphically and thereby provide intuition on the working mechanism of an envelope analysis. Figure 1.1 shows a plot of Y_6 versus Y_7 with the points marked by treatment. Since $\boldsymbol{\beta} = \text{E}(\mathbf{Y}|X=1) - \text{E}(\mathbf{Y}|X=0) \in \mathbb{R}^{2\times 1}$, the standard estimator for $\boldsymbol{\beta}$ is obtained as the difference in the marginal means after projecting the data onto the horizontal and vertical axes of the plot. The two densities estimates with the larger variation shown along the horizontal axes of the plot represent this operation. These density estimates are nearly identical, which explains the relatively small t-values from the standard model mentioned previously. However, it is clear from the figure that the treatments do differ.

An envelope analysis infers that $\boldsymbol{\beta} = (\beta_6, \beta_7)^T$ is parallel to the second eigenvector of $\boldsymbol{\Sigma}_{\mathbf{Y}|X} = \text{cov}(Y_6, Y_7)$. Hence by Proposition 1.1, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|X}}(\boldsymbol{\beta}) = \text{span}(\boldsymbol{\beta})$, as shown on the plot. The envelope represents the subspace in which the populations differ, which seems consistent with the pattern of variation shown in the plot. The orthogonal complement of the envelope, represented by a dashed line on the plot, represents the immaterial variation. The two populations are inferred to be the same when projected onto this subspace, which

also seems consistent with the pattern of variation in the plot. The envelope estimator of a mean difference is obtained by first projecting the points onto the envelope and thus removing the immaterial variation, and then projecting the points onto the horizontal or vertical axis. The two density estimates with the smaller variation represent this operation. These densities are well separated, leading to increased efficiency.

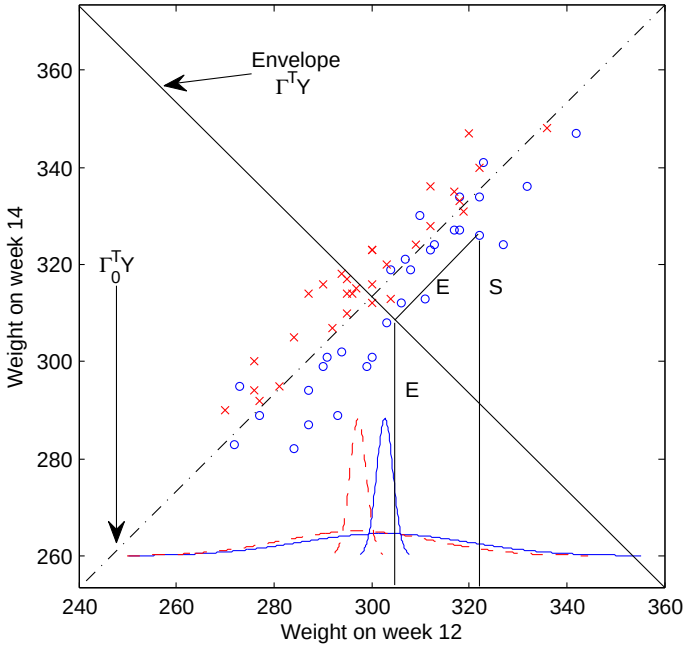


Figure 1.1: Kenward's cow data with the 30 animals receiving one treatment marked as o's and the 30 animals receiving the other marked as x's.The curves on the bottom are densities of $Y_6|(X = 0)$ and $Y_6|(X = 1)$: the flat two curves are obtained by projecting the data onto the Y_6 axis (standard analysis), and the two other densities are obtained by first project the data onto the envelope and then onto the Y_6 axis (envelope analysis). Representative projection paths, labeled 'E' for envelope analysis and 'S' for standard analysis, are shown on the plot.

OLS estimator										
Week	2	4	6	8	10	12	14	16	18	19
$\widehat{\boldsymbol{\beta}}_{\text{OLS}}$	2.4	3.3	3.1	4.7	4.7	5.5	-4.8	-4.5	-2.8	5.0
Bootstrap SE	2.9	3.2	3.5	3.6	4.0	4.2	4.4	4.5	5.4	6.0
Envelope estimator										
$\widehat{\boldsymbol{\beta}}_{\text{env}}$	-2.2	-0.5	0.9	2.4	2.9	5.4	-5.1	-4.6	-3.7	4.2
Bootstrap SE	1.13	0.84	1.07	1.03	0.81	1.12	1.07	1.04	1.08	1.02
Asymptotic SE/$\sqrt{n}$	0.88	0.74	0.72	0.84	0.70	1.02	0.92	0.86	0.90	0.85
Bootstrap SE ratios of OLS estimator over envelope estimator										
SE ratios	2.6	3.8	3.3	3.5	5.0	3.7	4.1	4.3	5.0	5.9

Table 1.1: Bootstrap standard errors of the 10 elements in $\widehat{\boldsymbol{\beta}}$ under the OLS estimator and the envelope estimator with $u = 1$. The bootstrap standard errors were estimated using 100 bootstrap samples.

Chapter 2

Algorithms for Envelope Estimation

2.1 Introduction

As mentioned earlier in Chapter 1, envelope methods aim to reduce estimative variation in multivariate linear models. The reduction is typically associated with predictors or responses, and can generally be interpreted as effective dimensionality reduction in the parameter space. Such reduction is achieved by enveloping the variation in the data that is material to the goals of the analysis while simultaneously excluding the immaterial variation. Efficiency gains are then achieved by essentially basing estimation on the material variation alone. The improvement in estimation and prediction can be quite substantial when the immaterial variation is large, sometimes equivalent to taking thousands of additional observations.

All current envelope methods, for example Cook et al. (2010; 2013), are all likelihood-based. The likelihood-based approach to envelope estimation requires, for a given envelope dimension u, optimizing an objective function of the form $f(\mathbf{\Gamma})$, where $\mathbf{\Gamma}$ is a $k \times u$, $k > u$, semi-orthogonal basis matrix for the envelope. The objective function satisfies $f(\mathbf{\Gamma}) = f(\mathbf{\Gamma O})$ for any $u \times u$ orthogonal matrix $\mathbf{O}$. Hence the optimization is essentially over the set of all u-dimensional subspaces of $\mathbb{R}^k$, which is a Grassmann manifold denoted as $\mathcal{G}_{u,k}$. Since $u(k-u)$ real numbers are required to specify an element of $\mathcal{G}_{u,k}$ uniquely, the optimization is essentially over $u(k-u)$ real dimensions. In multivariate linear regression, k can be either the number of responses r or the number of predictors p, depending on whether one is pursuing response or predictor reduction.

All present envelope methods rely on the Matlab package `sg_min` by Ross A. Lippert (`http://web.mit.edu/~ripper/www/software/`) to optimize $f(\mathbf{\Gamma})$. This package provides iterative optimization techniques on Stiefel and Grassmann manifolds, including

non-linear conjugate gradient (PRCG and FRCG) iterations, dog-leg steps and Newton's method. For more background in Grassmann manifolds and Grassmann optimizations, see Edelman, Tomas and Smith (1998) and Absil, Mahony and Sepulchre (2008). Since Grassmann manifold optimization is not popular in statistics, it is worth mentioning that there are two packages for sufficient dimension reduction methods using Grassmann manifold optimization: R package `GrassmannOptim` by Adragni, Cook and Wu (2012) and Matlab package `LDR` by Cook, Forzani and Tomassi (2011), which uses `sg_min` to implement several sufficient dimension reduction methods

To implement an envelope estimation procedure, one needs to specify the objective function $f(\mathbf{\Gamma})$ and its analytical first-order derivative function. Then given an initial value of $\mathbf{\Gamma}$, this package will compute numerical second-order derivatives and iterate until convergence or the maximum number of iterations is reached. The Matlab toolbox `envlp` by Cook, Su and Yang (`http://code.google.com/p/envlp/`) uses `sg_min` to implement a variety of envelope estimators along with associated inference methods. The `sg_min` package works well for envelope estimation, but nevertheless, optimization is often computationally difficult for large values of $u(k-u)$. At higher dimensions, each iteration becomes exponentially slower, local minima can become a serious issue and good starting values are essential. The `envlp` toolbox implements a seemingly different version of $f(\mathbf{\Gamma})$ for each type of envelope, along with tailored starting values.

In this chapter we present two advances in envelope computation. First, we propose in Section 2.2 a model-free objective function $J_n(\mathbf{\Gamma})$ for estimating an envelope and show that the three major envelope methods are based on special cases of J_n. This unifying objective function is to be optimized over the Grassmann manifold $\mathcal{G}_{u,k}$, which for larger values of $u(k-u)$ will be subject to the same computational limitations associated with speed, local minima and starting values. Second, we propose in Section 2.3 a fast one-dimensional (1D) algorithm that mitigates these computational issues. To adapt the envelope construction for relatively large values of $u(k-u)$, we break down Grassmann optimization into a series of one-dimensional optimizations so that the estimation procedure is speeded up greatly, and starting values and local minima are no longer an issue. Although it may be impossible to break down a general u-dimensional Grassmann optimization problem, we rely on special characteristics of envelopes in statistical problems to achieve the breakdown of envelope estimation. The resulting 1D algorithm, which is easy-to-implement, stable and requires no initial value input, can be tens to hundreds times faster than the general Grassmann manifold optimization for $u > 1$, while still providing a desirable $\sqrt{n}$-consistent envelope estimator. We will use special forms of the 1D algorithm to find initial values for the simultaneous envelope in Chapter 3 and for reduced-rank envelope in Chapter 4. The 1D algorithm we introduce in Section 2.3 is much more general than its use in those two chapters and is directly applicable beyond the multivariate linear regression context,

see Chapter 5. The rest of this chapter is organized as follows. Section 2.4 consists of simulation studies and a data example to further demonstrate the advantages of the 1D algorithm. Section 2.5 is a brief conclusion of this chapter. Proofs and technical details are included in the Section 2.6.

2.2 Objective functions for estimating an envelope

2.2.1 The objective function and its properties

In this section we propose a generic objective function for estimating a basis $\mathbf{\Gamma}$ of an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) \subseteq \mathbb{R}^d$, where $\mathbf{M} \in \mathbb{S}^d$ is a symmetric positive definite matrix. Let $\mathcal{B}$ be spanned by a $d \times d$ matrix $\mathbf{U}$ so that $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$. Because $\text{span}(\mathbf{U}) = \text{span}(\mathbf{U}\mathbf{U}^T)$, we can always denote the envelope by $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ for some symmetric matrix $\mathbf{U} \geq 0$. We propose the following generic population objective function for estimating $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$:

$$J(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\mathbf{M}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{\Gamma}|, \tag{2.2.1}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{d\times u}$ denotes a semi-orthogonal basis for elements in Grassmann manifold $\mathcal{G}_{u,d}$, u is the dimension of the envelope, and $u < d$. We refer to the operation of optimizing (2.2.1) or its sample version given later in (2.2.2) as full Grassmann (FG) optimization. Since $J(\mathbf{\Gamma}) = J(\mathbf{\Gamma}\mathbf{O})$ for any orthogonal $u \times u$ matrix $\mathbf{O}$, the minimizer $\widetilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} J(\mathbf{\Gamma})$ is not unique. But we are interested only in $\text{span}(\widetilde{\mathbf{\Gamma}})$, which is unique as shown in the following proposition.

Proposition 2.1. *Let $\widetilde{\mathbf{\Gamma}} \in \mathbb{R}^{d\times u}$ be a minimizer of $J(\mathbf{\Gamma})$. Then* $\text{span}(\widetilde{\mathbf{\Gamma}}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$.

To gain intuition on how $J(\mathbf{\Gamma})$ is minimized by any $\widetilde{\mathbf{\Gamma}}$ that spans the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$, we let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{d\times d}$ be an orthogonal matrix and decompose the objective function into two parts: $J(\mathbf{\Gamma}) = J^{(1)}(\mathbf{\Gamma}) + J^{(2)}(\mathbf{\Gamma})$, where

$$\begin{aligned} J^{(1)}(\mathbf{\Gamma}) &= \log|\mathbf{\Gamma}^T\mathbf{M}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0|, \\ J^{(2)}(\mathbf{\Gamma}) &= \log|\mathbf{\Gamma}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{\Gamma}| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0| \\ &= \log|\mathbf{\Gamma}_0^T(\mathbf{M}+\mathbf{U})\mathbf{\Gamma}_0| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0| - \log|\mathbf{M}+\mathbf{U}|. \end{aligned}$$

The first function $J^{(1)}(\mathbf{\Gamma})$ is minimized by any $\mathbf{\Gamma}$ that spans a reducing subspace of $\mathbf{M}$. Minimizing the second function $J^{(2)}(\mathbf{\Gamma})$ is equivalent to minimizing $\log|\mathbf{\Gamma}_0^T(\mathbf{M}+\mathbf{U})\mathbf{\Gamma}_0| - \log|\mathbf{\Gamma}_0^T\mathbf{M}\mathbf{\Gamma}_0|$, which is no less than zero and equals to zero when $\mathbf{\Gamma}_0^T\mathbf{U}\mathbf{\Gamma}_0 = 0$. Thus $J^{(2)}(\mathbf{\Gamma})$ is minimized by any $\mathbf{\Gamma}$ such that $\mathbf{\Gamma}_0^T\mathbf{U}\mathbf{\Gamma}_0 = 0$, or equivalently, $\text{span}(\mathbf{U}) \subseteq \text{span}(\mathbf{\Gamma})$. These properties of $J^{(1)}(\mathbf{\Gamma})$ and $J^{(2)}(\mathbf{\Gamma})$ are combined by $J(\mathbf{\Gamma})$ to get a reducing subspace of $\mathbf{M}$ that contains $\text{span}(\mathbf{U})$. In the context of multivariate linear regression, minimizing $J^{(2)}(\mathbf{\Gamma})$ is related to minimizing the residual sum of squares and minimizing $J^{(2)}$ is in effect pulling

the solution towards principal components of responses or predictors. Finally, because u is the dimension of the envelope, the minimizer $\text{span}(\widetilde{\mathbf{\Gamma}})$ is unique by Definition 1.2.

The sample version J_n of J based on a sample of size n is constructed by substituting estimators $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ of $\mathbf{M}$ and $\mathbf{U}$:

$$J_n(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\widehat{\mathbf{M}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{\Gamma}|. \tag{2.2.2}$$

Proposition 2.1 shows Fisher consistency of minimizers from optimizing the population objective function. Furthermore, $\sqrt{n}$-consistency of $\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} J_n(\mathbf{\Gamma})$ is stated in the following proposition.

Proposition 2.2. *Let $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ denote $\sqrt{n}$-consistent estimators for $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$. Let $\widehat{\mathbf{\Gamma}} \in \mathbb{R}^{d\times u}$ be a minimizer of $J_n(\mathbf{\Gamma})$, then $\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$.*

When we connect the objective function $J_n(\mathbf{\Gamma})$ with multivariate linear models in Section 2.2.2, we will find that previous likelihood-based envelope objective functions can be written in form (2.2.2). The likelihood approach to envelope estimation is based on normality assumptions for the conditional distribution of the response given the predictors or the joint distribution of the predictors and responses. The envelope objective function arising from this approach is a partially maximized log-likelihood obtained broadly a follows. After incorporating the envelope structure into the model, partially maximize the normal log-likelihood function $L_n(\boldsymbol{\psi}, \mathbf{\Gamma})$ over all the other parameters $\boldsymbol{\psi}$ with $\mathbf{\Gamma}$ fixed. This leads to a likelihood-based objective function $L_n(\mathbf{\Gamma})$, which equals a constant plus $-(n/2)J_n(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ depending on context. Proposition 2.2 indicates that the function $J_n(\mathbf{\Gamma})$ can be used as a generic moment-based objective function requiring only $\sqrt{n}$-consistent matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$. Consequently, normality is not a requirement for estimators based on $J_n(\mathbf{\Gamma})$ to be useful, a conclusion that is supported by previous work and by our experience. FG optimization of $J_n(\mathbf{\Gamma})$ can be computationally intensive and can require a good initial value. The 1D algorithm in Section 2.3 mitigates the computational issues.

2.2.2 Connections with previous work

Envelope applications have so far been mostly restricted to the homoscedastic multivariate linear model

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{2.2.3}$$

where $\mathbf{Y} \in \mathbb{R}^r$, the predictor vector $\mathbf{X} \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and the errors $\boldsymbol{\varepsilon}_i$ are independent copies of the normal random vector $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}})$. The maximum likelihood estimators of $\boldsymbol{\beta}$ and $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ are then $\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$.

Response envelopes

Cook, et al. (2010) studied response envelopes for estimation of the coefficient matrix $\boldsymbol{\beta}$. They conditioned on the observed values of $\mathbf{X}$ and motivated their developments by allowing for the possibility that some linear combinations of the response vector $\mathbf{Y}$ are immaterial to the estimation of $\boldsymbol{\beta}$, as described previously in Section 1.3.2. Reiterating, suppose that there is an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ so that (i) $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\boldsymbol{\Gamma})$ and (ii) $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$. This implies that $(\mathbf{Y}, \boldsymbol{\Gamma}^T\mathbf{X}) \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{X}$ and thus that $\boldsymbol{\Gamma}_0^T\mathbf{X}$ is immaterial to the estimation of $\boldsymbol{\beta}$. The smallest subspace $\text{span}(\boldsymbol{\Gamma})$ for which these conditions hold is the $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$-envelope of $\text{span}(\boldsymbol{\beta})$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$.

To determine the FG estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, we let $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$ in the objective function $J_n(\boldsymbol{\Gamma})$ to reproduce the likelihood-based objective function in Cook et al. (2010). Then the maximum likelihood envelope estimators are $\widehat{\boldsymbol{\beta}}_{\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X},\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$, where $\widehat{\boldsymbol{\Gamma}} = \arg\min J_n(\boldsymbol{\Gamma})$. Assuming normality for $\boldsymbol{\varepsilon}_i$, Cook et al. (2010) showed that the asymptotic variance of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is no larger than that of the usual least squares estimator $\widehat{\boldsymbol{\beta}}$. Under the weaker condition that $\boldsymbol{\varepsilon}_i$ are independent and identically distributed with finite fourth moments, the sample covariance matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent for $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{U} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\beta}^T$. By Proposition 2.2, we have $\sqrt{n}$-consistency of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ under this weaker condition.

Partial envelopes

Su and Cook (2011) used the $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$-envelope of $\text{span}(\boldsymbol{\beta}_1)$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, to develop a partial envelope estimator of $\boldsymbol{\beta}_1$ in the partitioned multivariate linear regression

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_1\mathbf{X}_{1i} + \boldsymbol{\beta}_2\mathbf{X}_{2i} + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{2.2.4}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{r\times p_1}$, $p_1 \le p$, is the parameter vector of interest, $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and the remaining terms are as defined for model (2.2.3). In this formulation, the immaterial information is $\boldsymbol{\Gamma}_0^T\mathbf{Y}$, where $\boldsymbol{\Gamma}_0$ is a basis for $\mathcal{E}^{\perp}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$. Since $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, the partial envelope estimator $\widehat{\boldsymbol{\beta}}_{1,\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}_1$ has the potential to yield efficiency gains beyond those for the full envelope, particularly when $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathbb{R}^r$ so the full envelope offers no gain. In the maximum likelihood estimation of $\boldsymbol{\Gamma}$, the same forms of $\widehat{\mathbf{M}}$, $\widehat{\mathbf{U}}$ and $J_n(\boldsymbol{\Gamma})$ are used for partial envelopes $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, except the roles of $\mathbf{Y}$ and $\mathbf{X}$ in the usual response envelopes are replaced with the residuals: $\mathbf{R}_{\mathbf{Y}|\mathbf{X}_2}$, residuals from the linear fits of $\mathbf{Y}$ on $\mathbf{X}_2$, and $\mathbf{R}_{\mathbf{X}_1|\mathbf{X}_2}$, the residuals of $\mathbf{X}_1$ on $\mathbf{X}_2$. Setting $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{R}_{\mathbf{Y}|\mathbf{X}_2}|\mathbf{R}_{\mathbf{X}_1|\mathbf{X}_2}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}$ in the objective function $J_n(\boldsymbol{\Gamma})$ reproduces the likelihood objective function of Su and Cook. Again, Proposition 2.2 gives $\sqrt{n}$-consistency without normality.

Predictor envelopes

Cook, et al. (2013) studied predictor reduction in model (2.2.3), except the predictors are now stochastic with $\text{var}(\mathbf{X}) = \mathbf{\Sigma_X}$ and $(\mathbf{Y}, \mathbf{X})$ was assumed to be normally distributed for the construction of maximum likelihood estimators. Their reasoning, which parallels that for response envelopes, lead them to parameterize the linear model in terms of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$ and to achieve similar substantial gains in the estimation of $\boldsymbol{\beta}$ and in prediction. The immaterial information in this setting is given by $\mathbf{\Gamma}_0^T\mathbf{X}$, where $\mathbf{\Gamma}_0$ is now a basis for $\mathcal{E}^{\perp}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$. They also showed that the SIMPLS algorithm for partial least squares provides a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$ and demonstrated that the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\beta}}\mathbf{P}^T_{\widehat{\mathbf{\Gamma}}(\mathbf{S_X})}$ typically outperforms the SIMPLS estimator in practice. For predictor reduction in model (2.2.3), the envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$ is estimated with $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{X}|\mathbf{Y}}$, $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S_X}$. As with response and partial envelopes, Proposition 2.2 gives us $\sqrt{n}$-consistency without requiring normality for $(\mathbf{Y}, \mathbf{X})$.

Techniques for estimating the dimension of an envelope are discussed in the parent articles of these methods, including use of an information criterion like BIC, cross validation or a hold-out sample.

2.2.3 New envelope estimators inspired by the objective function

The objective function $J_n(\mathbf{\Gamma})$ can also be used for envelope estimation in new problems. For example, to estimate the multivariate mean $\boldsymbol{\mu} \in \mathbb{R}^r$ in the model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, we can use the $\mathbf{\Sigma_Y}$-envelope of $\text{span}(\boldsymbol{\mu})$ by taking $\mathbf{M} = \mathbf{\Sigma_Y}$ and $\mathbf{U} = \boldsymbol{\mu}\boldsymbol{\mu}^T$, whose sample versions are: $\widehat{\mathbf{M}} = \mathbf{S_Y}$, $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T$ and $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \mathbf{Y}_i$. Then substituting $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ leads to the same objective function $J_n(\mathbf{\Gamma})$ as that obtained when deriving the likelihood-based envelope estimator from scratch.

For the second example, let $\mathbf{Y}_i \sim N_r(\boldsymbol{\mu}, \mathbf{\Sigma_Y})$, $i = 1, \ldots, n$, consist of longitudinal measurements of n subjects over r fixed time points. Suppose we are not interested in the overall mean $\overline{\mu} = \mathbf{1}_r^T\boldsymbol{\mu}/r \in \mathbb{R}^1$ but rather interest centers on the deviations at each time point $\boldsymbol{\alpha} = \boldsymbol{\mu} - \overline{\mu}\mathbf{1}_r \in \mathbb{R}^r$. Let $\mathbf{Q_1} = \mathbf{I}_r - \mathbf{1}_r\mathbf{1}_r^T/r$ denote the projection onto the orthogonal complement of $\text{span}(\mathbf{1}_r)$. Then $\boldsymbol{\alpha} = \mathbf{Q_1}\boldsymbol{\mu}$ and we consider estimating the constrained envelope: $\mathcal{E}_{\mathbf{Q_1}\mathbf{\Sigma_Y}\mathbf{Q_1}}(\mathbf{Q_1}\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{Q_1}) := \mathcal{E}_{\mathbf{M}}(\mathbf{U})$. Optimizing $J_n(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}} = \mathbf{Q_1}\mathbf{S_Y}\mathbf{Q_1}$ and $\widehat{\mathbf{U}} = \mathbf{Q_1}\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T\mathbf{Q_1}$ will again lead to the maximum likelihood estimator and to $\sqrt{n}$-consistency without normality. Later from Proposition 2.3, we will see that $\mathcal{E}_{\mathbf{M}}(\mathbf{U}) = \mathbf{Q_1}\mathcal{E}_{\mathbf{\Sigma_Y}}(\boldsymbol{\mu}\boldsymbol{\mu}^T)$ and the optimization can be simplified.

The objective function $J_n(\mathbf{\Gamma})$ introduces also a way of extending envelope regression semi-parametrically or non-parametrically. This can be done by simply replacing the sample covariances $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ in Section 2.2.2 with their semi-parametric and non-parametric counterparts. Given a multivariate model $\mathbf{Y} = \mathbf{f}(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{Y} \in \mathbb{R}^r$ and

$\mathbf{f}(\cdot) : \mathbb{R}^p \to \mathbb{R}^r$, the envelope for reducing the response can be estimated by taking $\widehat{\mathbf{M}}$ equal to the sample covariance of the residuals: $\widehat{\mathbf{M}} = n^{-1}\sum_{i=1}^{n}\{\mathbf{Y}_i - \widehat{\mathbf{f}}(\mathbf{X}_i)\}\{\mathbf{Y}_i - \widehat{\mathbf{f}}(\mathbf{X}_i)\}^T$, and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}}$.

2.3 A 1D algorithm

In this section we propose a method for estimating a basis $\boldsymbol{\Gamma}$ of an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) \subseteq \mathbb{R}^d$ based on a series of one-dimensional optimizations. The resulting algorithm is fast and stable, does not require carefully chosen starting values and the estimator it produces converges at the root-n rate. The estimator can be used as it stands, or as a $\sqrt{n}$-consistent starting value for (2.2.2). In the latter case, one Newton-Raphson step from the starting value provides an estimator that is asymptotically equivalent under normality to the maximum likelihood estimators discussed in Section 2.2.2 (Lehmann and Casella, 1998, p. 454.) The algorithm we present here can also be used for finding initial values of simultaneous envelope objective function in Chapter 3 and initial values of reduced-rank envelope objective function in Chapter 4.

The population algorithm described in this section extracts one dimension at a time from $\mathcal{E}_{\mathbf{M}}(\mathcal{B}) = \mathcal{E}_{\mathbf{M}}(\mathbf{U})$ until a basis is obtained. It requires only $\mathbf{M} > 0$, $\mathbf{U} \geq 0$ and $u = \dim(\mathcal{E}_{\mathbf{M}}(\mathcal{B}))$ as previously defined in Section 2.2. Sample versions are obtained by substituting $\sqrt{n}$-consistent estimators $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ for $\mathbf{M}$ and $\mathbf{U}$. Otherwise, the algorithm itself does not depend on a statistical context, although the manner in which the estimated basis is used subsequently does.

The following proposition is the basis for a sequential breakdown of a u-dimensional FG optimization.

Proposition 2.3. *Let* $(\mathbf{B}, \mathbf{B}_0)$ *denote an orthogonal basis of* $\mathbb{R}^d$*, where* $\mathbf{B} \in \mathbb{R}^{d\times q}$*,* $\mathbf{B}_0 \in \mathbb{R}^{d\times(d-q)}$ *and* $\operatorname{span}(\mathbf{B}) \subseteq \mathcal{E}_{\mathbf{M}}(\mathcal{B})$*. Then* $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ *implies that* $\mathbf{B}_0\mathbf{v} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$*.*

Suppose we know an orthogonal basis $\mathbf{B}$ for a subspace of the envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$. Then by Proposition 2.3 we can find the rest of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ by looking into $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$, which is a lower dimensional envelope. This then provides a motivation for Algorithm 1, which sequentially constructs vectors $\mathbf{g}_k \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$, $k = 1, \ldots, u$, until a basis is obtained, $\operatorname{span}(\mathbf{g}_1, \ldots, \mathbf{g}_u) = \mathcal{E}_{\mathbf{M}}(\mathcal{B})$. This algorithm follows the structure implied by Proposition 2.3 and the stepwise objective functions D_k are each one-dimensional versions of (2.2.1). The first direction $\mathbf{g}_1$ requires optimization in $\mathbb{R}^d$, while the optimization dimension is reduced by 1 in each subsequent step.

Remark 1. At step 2(c) of Algorithm 1, we need to minimize the stepwise objective function $D_k(\mathbf{w})$ under the constraint that $\mathbf{w}^T\mathbf{w} = 1$. The `sg_min` package can still be used to deal with this constraint since we are optimizing over one-dimensional Grassmann

Algorithm 1 The 1D algorithm.

1. Set initial value $\mathbf{g}_0 = \mathbf{G}_0 = 0$.
2. For $k = 0, \ldots, u-1$,
   (a) Let $\mathbf{G}_k = (\mathbf{g}_1, \ldots, \mathbf{g}_k)$ if $k \geq 1$ and let $(\mathbf{G}_k, \mathbf{G}_{0k})$ be an orthogonal basis for $\mathbb{R}^d$.
   (b) Define the stepwise objective function

$$D_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{M}_k \mathbf{w}) + \log\{\mathbf{w}^T (\mathbf{M}_k + \mathbf{U}_k)^{-1} \mathbf{w}\}, \tag{2.3.1}$$

   where $\mathbf{M}_k = \mathbf{G}_{0k}^T \mathbf{M} \mathbf{G}_{0k}$, $\mathbf{U}_k = \mathbf{G}_{0k}^T \mathbf{U} \mathbf{G}_{0k}$ and $\mathbf{w} \in \mathbb{R}^{d-k}$.
   (c) Solve $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}} \mathrm{J}_k(\mathbf{w})$ subject to a length constraint $\mathbf{w}^T \mathbf{w} = 1$.
   (d) Define $\mathbf{g}_{k+1} = \mathbf{G}_{0k} \mathbf{w}_{k+1}$ to be the unit length $(k+1)$-th stepwise direction.

manifolds. An alternative way is to integrate the constraint $\mathbf{w}^T \mathbf{w} = 1$ into the objective function in (2.3.1), so that we only need to minimize the unconstrained function

$$\widetilde{D}_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{M}_k \mathbf{w}) + \log\{\mathbf{w}^T (\mathbf{M}_k + \mathbf{U}_k)^{-1} \mathbf{w}\} - 2\log(\mathbf{w}^T \mathbf{w}), \tag{2.3.2}$$

with an additional normalization step for its minimizer $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_{k+1}/||\mathbf{w}_{k+1}||$. This unconstrained objective function $D_k(\mathbf{w})$ can be solved by any standard numerical methods such as conjugate gradient or Newton's method. We have implemented this idea with the general purpose optimization function `optim` in R and obtained good results.

Remark 2. We have also considered other types of sequential optimization methods for envelope estimation. For example, we considered minimizing $D_1(\mathbf{w})$ at each step under orthogonality constraints such as $\mathbf{w}_{k+1}^T \mathbf{w}_j = 0$ or $\mathbf{w}_{k+1}^T \mathbf{M} \mathbf{w}_j = 0$ for $j \leq k$. These types of orthogonality constraints are used widely in PLS algorithms and principal components analysis. We find the statistical properties of these sequential methods are inferior to those of the 1D algorithm. For instance, they are clearly inferior in simulations and we doubt that they lead to consistent estimators.

The next two propositions establish the Fisher consistency of Algorithm 1 in the population and the $\sqrt{n}$-consistency of its sample version.

Proposition 2.4. *Assume that* $\mathbf{M} > 0$*, and let* $\mathbf{G}_u$ *denote the end result of the algorithm. Then* $\mathrm{span}(\mathbf{G}_u) = \mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

Proposition 2.5. *Assume that* $\mathbf{M} > 0$ *and let* $\widehat{\mathbf{M}} > 0$ *and* $\widehat{\mathbf{U}}$ *denote* $\sqrt{n}$*-consistent estimators for* $\mathbf{M}$ *and* $\mathbf{U}$*. Let* $\widehat{\mathbf{G}}_u$ *denote the estimator obtained from the 1D algorithm using* $\widehat{\mathbf{M}}$ *and* $\widehat{\mathbf{U}}$ *instead of* $\mathbf{M}$ *and* $\mathbf{U}$*. Then* $\mathbf{P}_{\widehat{\mathbf{G}}_u}$ *is* $\sqrt{n}$*-consistent for the projection onto* $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

The algorithm discussed in this section can be used straightforwardly in the contexts of the three envelopes reviewed in Section 2.2.2 and the extensions sketched in Section 2.2.3. The statistical properties of the 1D algorithm estimator stated in Propositions 2.4 and 2.5 are exactly parallel to the properties of FG optimization in Propositions 2.1 and 2.2.

2.4 Simulations

In this section, we compare the 1D algorithm to FG (full Grassmann manifold) optimization, focusing on computational cost. For fair comparisons, the implementation of our 1D algorithm was based on minimizing the length-constrained objective function (2.3.1) using the `sg_min` package. Implementation of the 1D algorithm with other computing packages using the unconstrained objective function (2.3.2) may offer even faster estimation procedures.

2.4.1 Simulations

We considered the response envelope model in Cook et al. (2010) with univariate predictor $X \sim N(0,1)$ and multivariate response $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}X + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_r(0, \boldsymbol{\Sigma}_{\mathbf{Y}|X})$ and we were interested in estimation of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|X}}(\boldsymbol{\beta})$. We generated $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{Y}|X}$ and $\mathbf{U} = \boldsymbol{\beta}\boldsymbol{\beta}^T$ in accordance with an envelope structure: $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ for some positive definite matrices $\boldsymbol{\Omega} \in \mathbb{S}^u$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{r-u}$ and a vector of ones $\boldsymbol{\eta} = \mathbf{1}_u \in \mathbb{R}^u$. The semi-orthogonal basis $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ for $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$ was randomly generated and $\boldsymbol{\Gamma}_0$ was then obtained so that $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was an orthogonal basis for $\mathbb{R}^r$. The two covariance matrices $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ were generated as $\mathbf{A}\mathbf{A}^T > 0$, where $\mathbf{A}$ was a square matrix with corresponding dimensions and was filled with uniform $(0,1)$ random numbers.

We first examined the performances of our 1D algorithm in the population. We generated 100 pairs of $\mathbf{M}$ and $\mathbf{U}$ for each of three dimension configurations, $(r,u) = (10,3)$, $(r,u) = (30,10)$ and $(r,u) = (70,20)$. These dimensions correspond to the real optimization dimensions $u(r-u) = 21$, 200 and 1000 for FG optimization, while the 1D algorithm optimizes over at most $r-1$ real dimensions at each iteration. We recorded the CPU time T for estimating an envelope and the Frobenius norm between the true envelope and an estimated envelope defined as $\text{dist}(\Gamma, \widetilde{\Gamma}) = ||\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T - \widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Gamma}}^T||_F$. The results for running the 1D algorithm (Algorithm 1) and the FG optimization of (2.2.2) are given in the first three rows of Table 2.1. Apparently the 1D algorithm achieved the same accuracy as FG optimization and was much less time-consuming, especially at the large dimension $(r,u) = (30,10)$ and $(r,u) = (70,20)$.

We next generated 100 replicated data sets for one pairs of $\mathbf{M}$ and $\mathbf{U}$, and used the sample estimator $\widehat{\mathbf{M}} = \mathbf{S}_{\mathbf{Y}|X}$ and $\widehat{\mathbf{U}} = \mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{Y}|X}$ for envelope estimation. We let $n = 400$ and kept the same dimensions. From Table 2.1, we can see the 1D algorithm outperformed

FG optimization in terms of computational efficiency.

For FG optimization, we chose initial value according to the approach described in Su and Cook (2011; Section 3.5), first optimizing the objective function over the $2r$ eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. This initial value search procedure alone could be computationally costly, but we did not include the time spent on this when we summarized the computing time T for the FG optimization algorithm in Table 2.1. Additionally, we used only the true value of u in each simulation. The performance of optimizations at other than the true value of u, as necessary in the application of BIC, need not follow those of Table 2.1, as we illustrate in the next section.

	1D algorithm		FG optimization	
	T	$\text{dist}(\boldsymbol{\Gamma}, \widehat{\boldsymbol{\Gamma}})$	T	$\text{dist}(\boldsymbol{\Gamma}, \widehat{\boldsymbol{\Gamma}})$
$(n, r, u) = (\infty, 10, 3)$	2.0 (0.2)	$< 1.0 \times 10^{-8}$	6.6 (0.3)	$< 1.0 \times 10^{-8}$
$(n, r, u) = (\infty, 30, 10)$	2.6 (0.1)	$< 1.0 \times 10^{-4}$	127 (11)	$< 1.0 \times 10^{-4}$
$(n, r, u) = (\infty, 70, 20)$	447 (11)	$< 1.0 \times 10^{-2}$	5084 (1283)	$< 1.0 \times 10^{-2}$
$(n, r, u) = (400, 10, 3)$	0.6 (0.04)	1.1 (0.05)	1.2 (0.09)	1.0 (0.05)
$(n, r, u) = (400, 30, 10)$	30.7 (0.6)	2.8 (0.02)	121 (7)	3.1 (0.02)
$(n, r, u) = (400, 70, 20)$	534 (5)	4.6 (0.04)	4187 (68)	4.7 (0.03)

Table 2.1: Comparisons between the 1D algorithm and FG optimization. Each cell contains the average running time in seconds over 100 simulations, with its standard error given in parentheses. The population algorithms with $\mathbf{M}$ and $\mathbf{U}$ were indicated with $n = \infty$ and the sample algorithms had $n = 400$.

2.4.2 Starting values

As mentioned previously, good starting values can be crucial to the performance of FG optimization. To highlight this point, we used the meat data analyzed previously by Cook et al. (2013) for envelope predictor reduction in multivariate linear regression. This data set consists of spectral measurements from infrared transmittance for fat, protein and water for 103 meat samples. Following Cook et al. (2013), we used the protein percentage as the univariate response. The $p = 50$ predictors were spectral measurements at every fourth wavelength between 850nm and 1050nm. Using five-fold cross-validation prediction error as their criterion and u varying from 1 to 25, Cook et al. (2013) compared the FG envelope estimator described in Section 2.2.2 to the OLS and SIMPLS estimators. The starting value for the FG envelope estimator was the SIMPLS estimator, which is $\sqrt{n}$-consistent in the context of predictor envelopes and had better performance than OLS. SIMPLS was designed specifically for predictor reduction and is not applicable to response or partial reduction or to the extensions discussed in Section 2.2.3. Their results showed the envelope estimator to be uniformly superior to OLS, superior to SIMPLS for small values of u and about the same as SIMPLS for large values of u. In this study we used

the same setup as Cook et al. (2013), except we focused on comparisons between the 1D algorithm and the FG envelope estimator with starting values again chosen following the approach described in Su and Cook (2011; Section 3.5), since the $2r$ eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$ may be all that is easily available without recourse to the 1D algorithm.

We plotted in Figure 2.1 (top two plots) the five-fold cross-validation squared prediction error and the elapsed CPU time (in seconds) for computing the FG envelope estimators with dimensions $u = 1, \ldots, 25$. Although we had five-folds and thus estimated five envelopes for each dimension u, the time reported is the average for estimating one envelope. The number of real optimization dimensions $u(50 - u)$ varied between 49 and 625. For the larger values of u, FG optimization took a very long time to compute, so we capped the number of allowed iterations at 5000. For small dimensions, $u \leq 3$, FG optimization and the 1D algorithm had close prediction performance, and there were no convergence issues. For $u = 4$ and 5, FG optimizations tended to become trapped into local minima, as indicated by the prediction error. For larger dimensions, $u > 10$, FG optimization began bumping into the iteration limit. The computation time for the 1D method was almost linearly increasing in u because of the sequential manner of the algorithm. With increasing number of components, the prediction errors of both methods converged towards that of the ordinary least squares estimator as expected, since they both reduce to ordinary least squares when $u = 50$. However, the 1D algorithm provided better estimators, consistently over u, than the OLS estimator and the FG envelope estimator.

This difference in the results reported by Cook et al. (2013) and the results shown in the top plot of Figure 2.1 arises because of the different staring values. In Cook et al. (2013), the initial values were $\sqrt{n}$-consistent, while here we chose initial values from the eigenvectors of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}} + \widehat{\mathbf{U}}$. When using these starting values, FG optimizations tended to get trapped by local minima that were close to the initial values, which accounts for the inferior performance of the FG envelope estimator in this setting. From Lehmann and Casella (1998; Theorem 4.3), we know that one Newton-Raphson iteration from any $\sqrt{n}$-consistent estimator, the 1D algorithm estimator for instance, will be asymptotically equivalent to the MLE, even if there were local minima. We used 100 iterations (instead of one) for the FG optimization with 1D algorithm estimators as initial values. The cross-validation prediction errors, shown in the bottom plot of Figure 2.1, were very close to those of the 1D algorithm. The FG algorithm did a little bit worse than the 1D algorithm at some u because with 100 iterations it occasionally got trapped in a local minimum as it tried to improve the starting value.

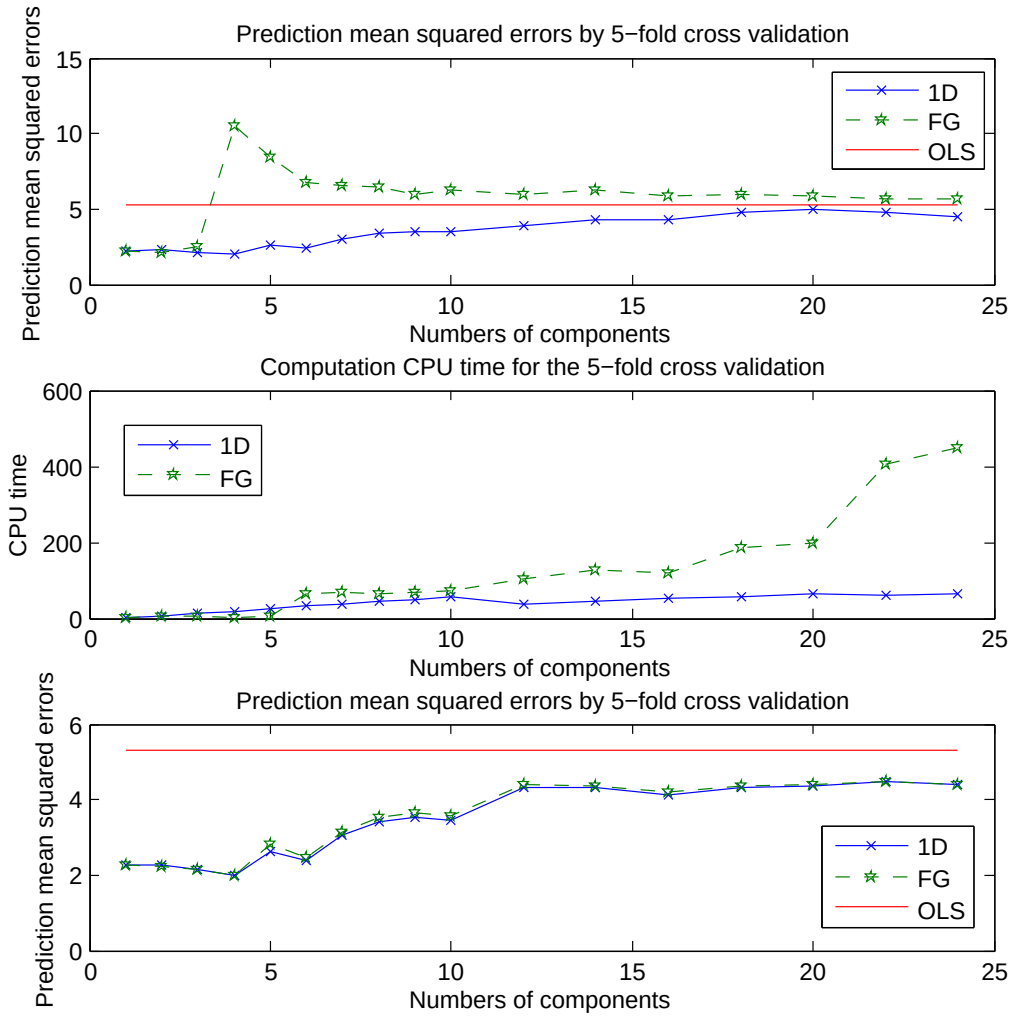


Figure 2.1: Meat protein data. The FG optimizations shown in the top and the middle plots were based on starting values suggested by Cook and Su (2011; Section 5.3). And the FG optimization in bottom plot was using the 1D algorithm estimators as starting value.

2.5 Conclusion

Our study led to the following conclusions. The FG envelope estimator (2.2.2) can be computed straightforwardly when the number of real dimensions $u(k-u)$ is relatively small, say less than 150, as illustrated in the example of Section 1.3.2. When this dimension is large, computing time and local minima can become serious issues, and then root-n consistent starting values become crucial. The 1D algorithm can be used confidently for starting values, or as a stand-alone algorithm for envelope estimation.

2.6 Proofs

2.6.1 Proposition 2.1

The proof of this proposition is very similar to the proof of Proposition 4.2 in Cook et al. (2013), thus is omitted.

2.6.2 Proposition 2.2

The proof follows from Proposition 2.6 and Proposition 2.7 in the same way as Proposition 2.5 in Section 2.6.5. Thus we omit the details of the proof.

2.6.3 Proposition 2.3

Proof. From our set-up, we know that $\mathbf{B}^T\mathbf{M}\mathbf{B} > 0$ thus $\mathcal{E}_{\mathbf{B}^T\mathbf{M}\mathbf{B}}(\mathbf{B}^T\mathcal{B})$ exists. Let $\mathbf{\Gamma}$ be a basis of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be a orthogonal basis of $\mathbb{R}^p$, then $\mathbf{M} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and $\mathcal{B} \subseteq \text{span}(\mathbf{\Gamma})$ for some symmetric matrices $\mathbf{\Omega} > 0$ and $\mathbf{\Omega}_0 > 0$. Therefore,

$$\begin{aligned} \mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 &= (\mathbf{B}_0^T\mathbf{\Gamma})\mathbf{\Omega}(\mathbf{B}_0^T\mathbf{\Gamma})^T + (\mathbf{B}_0^T\mathbf{\Gamma}_0)\mathbf{\Omega}_0(\mathbf{B}_0^T\mathbf{\Gamma}_0)^T \\ \mathbf{B}_0^T\mathcal{B} &\subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma}), \end{aligned} \tag{2.6.1}$$

where $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is the orthogonal compliment of $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma}_0)$ in $\mathbb{R}^{p-q}$ since $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathbf{\Gamma})$. Then we see that

$$\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 = \mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}}, \tag{2.6.2}$$

which implies that $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is a reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ which also contains $\mathbf{B}_0^T\mathcal{B}$ by (2.6.1). By definition, we know that $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ is the smallest reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ that contains $\mathbf{B}_0^T\mathcal{B}$. Hence $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B}) \subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$. Thus $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{B})$ implies $\mathbf{B}_0\mathbf{v} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$.

□

2.6.4 Proposition 2.4

Proof. We first write

$$\begin{aligned}\mathbf{M} &= \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T,\\ \mathbf{M}+\mathbf{U} &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T,\end{aligned}$$

where $\mathbf{\Omega}_0 > 0$, $\mathbf{\Omega} > 0$, $\mathbf{\Phi} > 0$, $\mathbf{\Omega} - \mathbf{\Phi} \geq 0$, $\mathbf{\Gamma}$ is semi-orthogonal basis for $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^p$ is orthogonal basis for $\mathbb{R}^p$.

We begin by considering optimization for the first direction $\mathbf{g}_1 = \arg\min_{\mathbf{g}\in\mathbb{R}^p} D_0(\mathbf{g})$, where $D_0(\mathbf{g}) = \log(\mathbf{g}^T\mathbf{M}\mathbf{g}) + \log\{\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}\}$ and the minimization is subject to the constraint $\mathbf{g}^T\mathbf{g} = 1$. Let $\mathbf{g} = \mathbf{\Gamma}\mathbf{h} + \mathbf{\Gamma}_0\mathbf{h}_0$ for some $\mathbf{h} \in \mathbb{R}^u$ and $\mathbf{h}_0 \in \mathbb{R}^{(p-u)}$. Consider the optimization problem as the unconstrained problem,

$$\mathbf{g}_1 = \arg\min_{\mathbf{g}\in\mathbb{R}^p} \left\{\log(\mathbf{g}^T\mathbf{M}\mathbf{g}) + \log\{\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}\} - 2\log(\mathbf{g}^T\mathbf{g})\right\}.$$

Then we will have the same solution as the original problem up to an arbitrary scaling constant. Next, we plug-in these expressions for $\mathbf{g}$, $\mathbf{M}+\mathbf{U}$ and $\mathbf{M}$,

$$\begin{aligned}& \log(\mathbf{g}^T\mathbf{M}\mathbf{g}) + \log\{\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}\} - 2\log(\mathbf{g}^T\mathbf{g})\\ =\ & \log\{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0\} + \log\{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0\} - 2\log\{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0\}\\ \equiv\ & f(\mathbf{h}, \mathbf{h}_0).\end{aligned}$$

Taking partial derivative with respect to $\mathbf{h}_0$, we have

$$\frac{\partial}{\partial\mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) = \frac{2\mathbf{\Omega}_0\mathbf{h}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}\mathbf{h}_0}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} - \frac{4\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}.$$

To get local minimums we need to set $\frac{\partial}{\partial\mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) = 0$ which gives the following equality.

$$\left\{\frac{2\mathbf{\Omega}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0}\right\}\mathbf{h}_0 = \left\{\frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\right\}\mathbf{h}_0.$$

Define

$$\mathbf{A}_0 = \left\{\frac{2\mathbf{\Omega}_0}{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0}\right\} \Big/ \left\{\frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\right\}. \quad (2.6.3)$$

Since $\mathbf{\Omega}_0 > 0$, we know $\mathbf{A}_0 > 0$. Then $\mathbf{A}_0\mathbf{h}_0 = \mathbf{h}_0$ has solutions only as eigenvectors of $\mathbf{A}_0$. The eigenvectors of $\mathbf{A}_0$ are the same as those of $\mathbf{\Omega}_0$. Hence, $\mathbf{h}_0$ equals 0 or any eigenvector $\ell_k(\mathbf{\Omega}_0)$ of $\mathbf{\Omega}_0$. Therefore, the minimum value of $f(\mathbf{h}, \mathbf{h}_0)$ has to be obtained by 0 or $\ell_k(\mathbf{\Omega}_0)$ (since $\mathbf{h}_0 = \infty$ can be easily eliminated). If $\mathbf{h}_0 = 0$ then our conclusion follows.

Assume $\mathbf{h}_0 \neq 0$ and $\mathbf{\Omega}_0\mathbf{h}_0 = \lambda_k\mathbf{h}_0$. Then,

$$\begin{aligned} f(\mathbf{h}, \mathbf{h}_0) &= \log\{\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h} + \lambda_k\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \frac{1}{\lambda_k}\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\}\\ &= \log\{\frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \lambda_k(1-W_h)\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \frac{1}{\lambda_k}(1-W_h)\},\end{aligned}$$

where $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0}$ is the weight between 0 and 1. Because log() is concave, we have $\log(aW_h + b(1-W_h)) \geq W_h \log(a) + (1-W_h)\log(b)$. Hence,

$$\begin{aligned} f(\mathbf{h},\mathbf{h}_0) &\geq W_h \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} + (1-W_h) \left\{ \log(\lambda_k) + \log(\frac{1}{\lambda_k}) \right\} \\ &= W_h \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \\ &\geq W_h \cdot \min_{\mathbf{h}\in\mathbb{R}^d} \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \\ &\geq \min_{\mathbf{h}\in\mathbb{R}^d} \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\}. \end{aligned}$$

The last inequality holds because

$$\min_{\mathbf{h}\in\mathbb{R}^d} \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} < 0, \tag{2.6.4}$$

which is proved in Section 2.6.4.

Moreover, the lower bound of $f(\mathbf{h},\mathbf{h}_0)$, which is negative, will be attained if we let $W_h = 1$ and let $\mathbf{h} = \arg\min_{\mathbf{h}\in\mathbb{R}^d} \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\}$. So we have the minimum found at $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0} = 1$, or equivalently, $\mathbf{g} = \mathbf{\Gamma}\mathbf{h} \in \text{span}(\mathbf{\Gamma})$.

For the $(k+1)$-th direction, $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1}$ where $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^{p-k}} D_k(\mathbf{w})$, subject to $\mathbf{w}^T\mathbf{w} = 1$. Because $D_k(\mathbf{w}) = \log(\mathbf{w}^T\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}\mathbf{w}) + \log[\mathbf{w}^T \left\{ \mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k} \right\}^{-1} \mathbf{w}]$ has the same form as $f(\mathbf{g})$, analogous to the first direction, this gives $\mathbf{w}_{k+1} \in \mathcal{E}_{\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}}(\mathbf{G}_{0k}^T\mathcal{B})$. Therefore $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1} \in \mathcal{E}_{\mathbf{M}}(\mathcal{B})$ by Proposition 2.3.

Proof of inequality (2.6.4)

We first show that $\min_{\mathbf{h}\in\mathbb{R}^d} \left\{ \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \leq 0$, then we assume the equality to conduct the proof by contradiction. Define the following two functions,

$$\begin{aligned} \mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1}) &:= \log \frac{\mathbf{h}^T\mathbf{\Phi}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}, \\ \mathrm{F}(\mathbf{h};\mathbf{\Omega},\mathbf{\Omega}^{-1}) &:= \log \frac{\mathbf{h}^T\mathbf{\Omega}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}, \end{aligned}$$

Recall that $\mathbf{\Omega} - \mathbf{\Phi} \geq 0$, hence $\mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h};\mathbf{\Omega},\mathbf{\Omega}^{-1})$ for any $\mathbf{h}$. Consider the minimum of both $\mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1})$ and $\mathrm{F}(\mathbf{h};\mathbf{\Omega},\mathbf{\Omega}^{-1})$, we have

$$\min_{\mathbf{h}} \mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1}) \leq \min_{\mathbf{h}} \mathrm{F}(\mathbf{h};\mathbf{\Omega},\mathbf{\Omega}^{-1}) = 0,$$

where the minimum of the right hand side is zero by taking $\mathbf{h}$ equals to any eigenvector of $\mathbf{\Omega}$.

Now we assume that $\min_{\mathbf{h}} \mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1}) = 0$. Then for an arbitrary $\mathbf{h}$,

$$0 \leq \mathrm{F}(\mathbf{h};\mathbf{\Phi},\mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h};\mathbf{\Omega},\mathbf{\Omega}^{-1}).$$

Let $\mathbf{h}_i = \ell_i(\mathbf{\Omega})$, $i = 1, \ldots, u$, be the i-th unit eigenvector of $\mathbf{\Omega}$ and plug $\mathbf{h}_i$ into the above inequalities, we have

$$0 \leq \mathrm{F}(\mathbf{h}_i; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) \leq \mathrm{F}(\mathbf{h}_i; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) = 0, \ i = 1, \ldots, u,$$

which implies

$$0 = \mathrm{F}(\mathbf{h}_i; \mathbf{\Phi}, \mathbf{\Omega}^{-1}) = \mathrm{F}(\mathbf{h}_i; \mathbf{\Omega}, \mathbf{\Omega}^{-1}) = 0, \ i = 1, \ldots, u,$$

and more explicitly,

$$\log(\mathbf{h}_i^T \mathbf{\Phi} \mathbf{h}_i) = \log(\mathbf{h}_i^T \mathbf{\Omega} \mathbf{h}_i), \ i = 1, \ldots, u,$$

which implies $\mathbf{\Phi} = \mathbf{\Omega}$ because that $\mathbf{\Phi}$, $\mathbf{\Omega} \in \mathbb{R}^{u \times u}$ and $\mathbf{h}_i$, $i = 1, \ldots, u$, are u linear independent vectors. Then by definition $\mathbf{U} = \mathbf{\Gamma}^T(\mathbf{\Phi} - \mathbf{\Omega})\mathbf{\Gamma} = 0$ leads to contradiction with the dimension of the envelope.

□

2.6.5 Proposition 2.5

Proof. Our proof of $\sqrt{n}$-consistency hinges on Amemiya's (1985) results on the asymptotic properties of extremum estimators. Proposition 4.1.1 and Proposition 4.1.3 in Amemiya (1985) can be applied to our context. We first state these results and then sketch how they can be used to prove the $\sqrt{n}$-consistency for our algorithm.

Let $Q_n(\mathbf{y}, \boldsymbol{\theta})$ be a real-valued function of the random variables $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ and the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^T$. We shall sometimes write $Q_n(\mathbf{y}, \boldsymbol{\theta})$ more compactly as $Q_n(\boldsymbol{\theta})$. Let the parameter space be $\mathbf{\Theta}$ and let the true value of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_t$ which is in $\mathbf{\Theta}$. Then Proposition 4.1.1 and Proposition 4.1.3 in Amemiya (1985) give asymptotic properties of the extremum estimator, $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \mathbf{\Theta}} Q_n(\mathbf{y}, \boldsymbol{\theta})$. We summarize the conditions in Amemiya's Propositions as follows.

(A) The parameter space $\mathbf{\Theta}$ is a compact subset of $\mathbb{R}^K$;

(B) $Q_n(\mathbf{y}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \mathbf{\Theta}$; for all $\mathbf{y}$ and is a measurable function of $\mathbf{y}$ for all $\boldsymbol{\theta} \in \mathbf{\Theta}$;

(C) $n^{-1}Q_n(\boldsymbol{\theta})$ converges to a nonstochastic function $Q(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta} \in \mathbf{\Theta}$ as n goes to infinity, and $Q(\boldsymbol{\theta})$ attains a unique global maximum at $\boldsymbol{\theta}_t$;

(D) $\partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ exists and is continuous in an open, convex neighborhood of $\boldsymbol{\theta}_0$;

(E) $n^{-1} \left\{ \partial^2 Q_n(\mathbf{y}, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \right\}_{\boldsymbol{\theta} = \boldsymbol{\theta}_n^*}$ converges to a finite nonsingular matrix

$$\mathbf{A}(\boldsymbol{\theta}_t) = \lim_{n \to \infty} \mathrm{E}_{\boldsymbol{\theta}_t} \left\{ n^{-1} \left\{ \partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \right\} \right\},$$

for any random sequences $\boldsymbol{\theta}_n^*$ such that $\mathrm{plim}(\boldsymbol{\theta}_n^*) = \boldsymbol{\theta}_t$;

(F) $n^{-1/2}\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \to N(0, \mathbf{B}(\boldsymbol{\theta}_t))$, where

$$\mathbf{B}(\boldsymbol{\theta}_t) = \lim_{n\to\infty} \mathrm{E}_{\boldsymbol{\theta}_t}\left\{n^{-1}\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}\right\}.$$

Proposition 2.6. *Under assumptions (A)-(C), $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_t$ in probability.*

Proposition 2.7. *Under assumptions (A)-(F), $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_t) \to N(0, \mathbf{A}(\boldsymbol{\theta}_t)^{-1}\mathbf{B}(\boldsymbol{\theta}_t)\mathbf{A}(\boldsymbol{\theta}_t)^{-1})$.*

In our adaptation of Proposition 2.6 and Proposition 2.7, we let $\boldsymbol{\theta} \equiv \mathbf{g}$ whose true value is denoted by $\mathbf{g}_t$ and let the random variables $\mathbf{y} = \mathrm{vech}(\widehat{\mathbf{M}}, \widehat{\mathbf{U}})$. The parameter space is the 1D manifold $\boldsymbol{\Theta} = \mathcal{G}_{(p,1)}$ which is a compact subset of $\mathbb{R}^p$, so condition (A) in Proposition 2.6 is satisfied. The function to be maximized is defined as follows.

$$Q_n(\mathbf{g}) = -n/2\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}) - n/2\log(\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}) + n\log(\mathbf{g}^T\mathbf{g}). \tag{2.6.5}$$

Condition (B) then holds. We next verify condition (C) that $n^{-1}Q_n(\mathbf{g})$ converges uniformly to

$$Q(\mathbf{g}) = -1/2\log(\mathbf{g}^T\mathbf{M}\mathbf{g}) - 1/2\log(\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}) + \log(\mathbf{g}^T\mathbf{g}). \tag{2.6.6}$$

We have shown that the population objective function $Q(\mathbf{g})$ attains the unique global maximum at $\mathbf{g}_t$. For simplicity, we assume $\mathbf{M}$ and $\mathbf{M}+\mathbf{U}$ both have distinct eigenvalues so that $\mathbf{g}_t$ is the unique maximum of $Q(\mathbf{g})$ in the 1D manifold $\boldsymbol{\Theta}$. For the case where there are multiple local maxima of $Q(\mathbf{g})$, we can obtain similar results by applying Proposition 4.1.2 in Amemiya (1985) as an alternative of Proposition 2.6. Since $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent for $\mathbf{M}$ and $\mathbf{U}$, the eigenvectors and eigenvalues of $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$ are $\sqrt{n}$-consistent for the eigenvectors and eigenvalues of their population counterparts.

Then $n^{-1}Q_n(\mathbf{g})$ converge in probability to $Q(\mathbf{g})$ uniformly in $\mathbf{g}$, as can be seen from the following argument.

$$\begin{aligned} n^{-1}Q_n(\mathbf{g}) - Q(\mathbf{g}) &= -1/2\left(\log(\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}) - \log(\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g})\right) \\ &\quad -1/2\left(\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}) - \log(\mathbf{g}^T\mathbf{M}\mathbf{g})\right) \\ &= -1/2\log\left[\frac{\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}}{\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}}\right] - 1/2\log\left[\frac{\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}}{\mathbf{g}^T\mathbf{M}\mathbf{g}}\right]. \end{aligned}$$

Hence, $\sup_{\mathbf{g}\in\boldsymbol{\Theta}}\log(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g}/\mathbf{g}^T\mathbf{M}\mathbf{g}) = \sup_{\mathbf{g}\in\boldsymbol{\Theta}}\log(\mathbf{g}^T\mathbf{M}^{-1/2}\widehat{\mathbf{M}}\mathbf{M}^{-1/2}\mathbf{g}/\mathbf{g}^T\mathbf{g})$, which equals to the logarithm of the largest eigenvalue of $\mathbf{M}^{-1/2}\widehat{\mathbf{M}}\mathbf{M}^{-1/2}$ and converges to 0 in probability. Similarly, $\sup_{\mathbf{g}\in\boldsymbol{\Theta}}\log[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}/\mathbf{g}^T(\mathbf{M}+\mathbf{U})^{-1}\mathbf{g}]$ converges to zero in probability. Therefore, $n^{-1}Q_n(\mathbf{g})$ converges to $Q(\mathbf{g})$ in probability uniformly in $\mathbf{g}\in\boldsymbol{\Theta}$. Note that we have assumed $\mathbf{M}+\mathbf{U} > 0$ and $\mathbf{M}^{-1} > 0$, so their eigenvalues will be bounded away from zero.

We next verify conditions $(D) - (F)$. By straightforward calculation, condition (D) follows from the second derivative matrix

$$\begin{aligned} n^{-1}\frac{\partial^2 Q_n(\mathbf{g})}{\partial \mathbf{g}\partial \mathbf{g}^T} &= 2(\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g})^{-2}(\widehat{\mathbf{M}}\mathbf{g}\mathbf{g}^T\widehat{\mathbf{M}}) - (\mathbf{g}^T\widehat{\mathbf{M}}\mathbf{g})^{-1}\widehat{\mathbf{M}} \\ &\quad +2\left[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\right]^{-2}\left[(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\right] \\ &\quad -\left[\mathbf{g}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}\right]^{-1}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1} \\ &\quad -2(\mathbf{g}^T\mathbf{g})^{-2}\mathbf{P}_{\mathbf{g}} + (\mathbf{g}^T\mathbf{g})^{-1}\mathbf{I}_p. \end{aligned} \tag{2.6.7}$$

Condition (E) holds because the above quantity is a smooth function of $\mathbf{g}$, $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$.

Last, we need to verify condition (F). From the proof of Proposition 2.7, we need only show that $n^{-1}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = O_p(1/\sqrt{n})$ for $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$. The derivative $n^{-1}\left\{\partial Q_n(\mathbf{g})/\partial\mathbf{g}\right\}_{\mathbf{g}=\mathbf{g}_t}$ equals

$$-(\mathbf{g}_t^T\widehat{\mathbf{M}}\mathbf{g}_t)^{-1}\widehat{\mathbf{M}}\mathbf{g}_t - (\mathbf{g}_t^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}_t)^{-1}(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}\mathbf{g}_t + 2\mathbf{g}_t. \tag{2.6.8}$$

Following the derivation for the population objective function, we know that

$$\left\{\partial Q(\mathbf{g})/\partial\mathbf{g}\right\}_{\mathbf{g}=\mathbf{g}_t} = 0.$$

Then the result follows from the fact that $n^{-1}\partial Q_n(\mathbf{g})/\partial\mathbf{g}$ is a smooth function of $\widehat{\mathbf{M}}$ and $(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})^{-1}$ which are $\sqrt{n}$-consistent estimators.

So far, we have verified the conditions $(A) - (F)$ so that the sample estimator $\widehat{\mathbf{g}}_1$ will be $\sqrt{n}$-consistent for the population estimator. For the $(k+1)$-th direction, $k < u$, let $\widehat{\mathbf{G}}_k$ denote an $\sqrt{n}$-consistent estimator of the first k directions and let $(\widehat{\mathbf{G}}_k, \widehat{\mathbf{G}}_{0k})$ be an orthogonal matrix. The $(k+1)$-th direction is defined by $\mathbf{g}_{k+1} = \widehat{\mathbf{G}}_{0k}\mathbf{w}_{k+1}$ where the parameters are $\mathbf{w}_{k+1} \in \boldsymbol{\Theta}_{k+1} \subset \mathbb{R}^{p-k}$ and the parameter space is $\boldsymbol{\Theta}_{k+1} = \mathcal{G}_{p-k,1}$. We show that we can obtain a $\sqrt{n}$-consistent estimator $\widehat{\mathbf{w}}_{k+1}$, so the $\sqrt{n}$-consistency of $\widehat{\mathbf{g}}_{k+1} = \widehat{\mathbf{G}}_{0k}\widehat{\mathbf{w}}_{k+1}$ then follows. We define our objective functions $Q_n(\mathbf{w})$ and $Q(\mathbf{w})$ as

$$\begin{aligned} Q_n(\mathbf{w}) &= -n/2\log(\mathbf{w}^T(\widehat{\mathbf{G}}_{0k}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\widehat{\mathbf{G}}_{0k})^{-1}\mathbf{w}) - n/2\log(\mathbf{w}^T\widehat{\mathbf{G}}_{0k}^T\widehat{\mathbf{M}}\widehat{\mathbf{G}}_{0k}\mathbf{w}) + n\log(\mathbf{w}^T\mathbf{w}) \\ Q(\mathbf{w}) &= -1/2\log(\mathbf{w}^T(\mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k})^{-1}\mathbf{w}) - 1/2\log(\mathbf{w}^T\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}\mathbf{w}) + \log(\mathbf{w}^T\mathbf{w}) \end{aligned}$$

Following the same logic as verifying the conditions for the first direction, we can see that $\widehat{\mathbf{w}} = \arg\max Q_n(\mathbf{w})$ will be $\sqrt{n}$-consistent for $\mathbf{v}_t = \arg\max Q(\mathbf{w})$ by noticing that $(\widehat{\mathbf{G}}_{0k}^T(\widehat{\mathbf{M}}+\widehat{\mathbf{U}})\widehat{\mathbf{G}}_{0k})^{-1}$ and $\widehat{\mathbf{G}}_{0k}^T\widehat{\mathbf{M}}\widehat{\mathbf{G}}_{0k}$ are $\sqrt{n}$-consistent estimators for $(\mathbf{G}_{0k}^T(\mathbf{M}+\mathbf{U})\mathbf{G}_{0k})^{-1}$ and $\mathbf{G}_{0k}^T\mathbf{M}\mathbf{G}_{0k}$. Since all the u directions will be $\sqrt{n}$-consistent, the projection onto $\widehat{\mathbf{G}}_u = (\widehat{\mathbf{g}}_1, \ldots, \widehat{\mathbf{g}}_u)$ will be a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$. □

Chapter 3

Simultaneous Envelopes for Multivariate Linear Regression

3.1 Introduction

Multivariate linear regression with predictors $\mathbf{X} \in \mathbb{R}^p$ and responses $\mathbf{Y} \in \mathbb{R}^r$ is a cornerstone of multivariate statistics. When p and r are not small, it is widely recognized that reducing the dimensionalities of $\mathbf{X}$ and $\mathbf{Y}$ may often result in improved performance. Perhaps the most popular methods for reducing the number of predictors and responses are principal component analysis (PCA), partial least squares (PLS) regression, canonical correlation analysis (CCA) and reduced-rank regression (RRR).

Principal component analysis (Jolliffe, 1986, 2005) is an un-supervised dimension reduction method designed to select orthogonal linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ with maximal variation. However, it does not use any information about the relationship between $\mathbf{X}$ and $\mathbf{Y}$, and thus separate PCA reductions of $\mathbf{X}$ and $\mathbf{Y}$ could be ineffective for regression. Partial least squares was proposed as iterative algorithms, NIPALS (Wold, 1966) and SIMPLS (de Jong, 1993), for predictor reduction. These methods, which are used extensively in chemometrics, reduce the predictors by iteratively estimating linear combinations of them that have maximal covariance with the response vector. Canonical correlation analysis (Hotelling, 1936; Anderson, 1984) is used to investigate the overall correlations between the two sets of variables $\mathbf{X}$ and $\mathbf{Y}$. It can simultaneously reduce $\mathbf{X}$ and $\mathbf{Y}$ by finding pairs of linear combinations such that the correlations of these pairs are in descending order and components are uncorrelated across different pairs. A probabilistic interpretation of CCA was given by Bach and Jordan (2005) as a latent variable model for two normal random vectors. Reduced-rank regression (Izenman 1975; Reinsel and Velu 1998) restricts on the rank of the regression coefficient matrix and therefore improves prediction by reducing the number of parameters in the model.

Sufficient dimension reduction methods for multivariate responses are also available for reducing dimensionality. For example, sliced inverse regression (Li, 1991) was extended to multivariate response data by Li, et al. (2003). Li, Wen and Zhu (2008) proposed "projective resampling" to deal with a multivariate regression by using univariate reduction methods. However, such methods are beyond the scope of this work since they are designed to estimate only a subspace as a preliminary step in an analysis. In contrast, we work in the context of the multivariate linear model with a view toward prediction and coefficient estimation.

Informally, multivariate linear regression can involve both material and immaterial variation in the responses and in the predictors. Material variation provides information that is directly relevant to the regression, while the immaterial variation is essentially irrelevant to the regression and serves to increase estimative variation. Envelopes, which were introduced by Cook, Li and Chiaromonte (2010) for response reduction, use a subspace to envelop the material information and thereby exclude the immaterial variation. Essentially a form of targeted dimension reduction, this process can lead to substantial efficiency gains when the immaterial variation is large relative to the material variation. Cook, Helland and Su (2013) adapted envelopes to the predictors, and showed that the SIMPLS algorithm for partial least squares regression converges to an envelope in the predictor space. Following Cook, et al. (2010), they demonstrated that using a likelihood-based objective function to separate the material and immaterial variation and to provide an estimator of the coefficient matrix produces clear and often substantial estimative and predictive advantages over SIMPLS.

However, little is known about using envelopes for joint reduction of the responses and predictors. The previous developments kindle a hope that we can combine their advantages to produce efficiency gains than are greater than those possible by reducing either the responses or the predictors alone. In this article, we develop likelihood-based envelope methods for simultaneously separating the material and immaterial variation in the responses and in the predictors. We show a potential for synergy in a synchronized reduction, producing an overall reduction in estimative variation surpassing that indicated by the marginal reductions. Finding a likelihood-based envelope can be computationally challenging, and so we propose a novel and fast optimization algorithm.

The rest of this chapter is organized as follows. We introduce our simultaneous reduction in Section 3.2, where we also link the simultaneous envelope method with partial least squares, canonical correlation analysis and reduced-rank regression. In Section 3.3, we derive a likelihood-based objective function that includes the objective functions used by Cook et al. (2010) and Cook et al. (2013) as special cases. Estimation procedure of a simultaneous envelope are also given in this section. In Section 3.4, asymptotic properties of the simultaneous envelope estimators are studied under normality and under general

distributional assumptions. Encouraging simulation results on prediction and on determine the dimensions of envelopes are given in Section3.6. In Section 3.7, we demonstrate the superiority of the simultaneous envelope estimator compared to classical methods by simulations and by predicting the contents of biscuit dough samples. Proofs and other technical details are included in Section 3.9.

3.2 Simultaneous envelopes

3.2.1 Definition and structure

The standard multivariate linear model can be written as

$$\mathbf{Y} = \boldsymbol{\mu}_Y + \boldsymbol{\beta}(\mathbf{X} - \boldsymbol{\mu}_X) + \boldsymbol{\epsilon}, \tag{3.2.1}$$

where $\boldsymbol{\mu}_Y$ is the mean for $\mathbf{Y}$, $\boldsymbol{\mu}_X$ is the mean for $\mathbf{X}$, $\boldsymbol{\epsilon}$ is the error vector that has mean 0, variance $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} > 0$ and is independent of $\mathbf{X}$, and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix in which we are primarily interested.

Envelope methods have the potential to increase efficiency in estimation of $\boldsymbol{\beta}$ after reducing $\mathbf{Y}$ (Cook et al., 2010) and to improve prediction of $\mathbf{Y}$ after reducing $\mathbf{X}$ (Cook et al., 2013). Our goal is to combine their advantages by simultaneously reducing $\mathbf{X}$ and $\mathbf{Y}$ to decrease both predictive and estimative variation. We next give a coordinate representation of simultaneous envelopes.

Let $d \leq \min(r, p)$ denote the rank of $\boldsymbol{\beta}$ and consider the singular value decomposition of $\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{r\times d}$ and $\mathbf{V} \in \mathbb{R}^{p\times d}$ are orthogonal matrices, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_d = \mathbf{V}^T\mathbf{V}$, and $\mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with elements $\lambda_1 \geq \cdots \geq \lambda_d > 0$ being the d singular values of $\boldsymbol{\beta}$. The column space (the left eigenspace) and the row space (the right eigenspace) of $\boldsymbol{\beta}$ are $\text{span}(\boldsymbol{\beta}) = \text{span}(\mathbf{U})$ and $\text{span}(\boldsymbol{\beta}) = \text{span}(\mathbf{V})$. Then two envelopes can be constructed for simultaneously reducing the predictor space and the response space:

1. $\mathbf{X}$-envelope: $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}^T)$ with $\dim(\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}^T)) = d_X$, $d \leq d_X \leq p$.

2. $\mathbf{Y}$-envelope: $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ with $\dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})) = d_Y$, $d \leq d_Y \leq r$.

We know from Cook, et al. (2010; Proposition 3.1) that $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$. Consequently, an alternative definition of the $\mathbf{Y}$-envelope is $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\boldsymbol{\beta})$, which then has the same form as $\mathbf{X}$-envelope, by replacing $\mathbf{X}$ with $\mathbf{Y}$ and $\text{span}(\boldsymbol{\beta}^T)$ with $\text{span}(\boldsymbol{\beta})$. We use $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ as the definition of the $\mathbf{Y}$-envelope to facilitate later parameterizations. If we imaging that the elements of $\boldsymbol{\beta}$ are generated with respect to Lebesque measure then it follows from Proposition 1.1 that no reduction is possible: $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}^T) = \mathbb{R}^p$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathbb{R}^r$ with probability one. Later in this section we show that proper envelopes imply certain relations between $\mathbf{X}$ and $\mathbf{Y}$ that may reasonably hold in practice.

From the definitions of the $\mathbf{X}$- and $\mathbf{Y}$-envelopes, if $d = r < p$ then we can reduce $\mathbf{X}$ only and $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathbb{R}^r$. Similarly, if $d = p < r$ then $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T) = \mathbb{R}^p$ and reduction is possible only in the response space. Hence, we will assume $d < \min(r,\ p)$ from now on and discuss the general situation where simultaneous reduction is possible. Let $\mathbf{R} \in \mathbb{R}^{p\times d_X}$ be an orthogonal basis for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and let $\mathbf{L} \in \mathbb{R}^{r\times d_Y}$ be an orthogonal basis for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$. Also, $(\mathbf{R}, \mathbf{R}_0)$ is an orthogonal basis for $\mathbb{R}^p$ and $(\mathbf{L}, \mathbf{L}_0)$ is an orthogonal basis for $\mathbb{R}^r$. Then from Definition 1.1 we can write the covariance matrices as

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{X}} &= \mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T, && (3.2.2)\\ \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}} &= \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T. && (3.2.3) \end{aligned}$$

The covariance matrix decomposition in (3.2.2) indicates that the eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ fall in either $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ or $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ with corresponding eigenvalues being the eigenvalues of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. No relationship is assumed between the eigenvalues of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$: The eigenvalues of $\boldsymbol{\Omega}$ could be any subset of the eigenvalues of $\mathbf{\Sigma}_{\mathbf{X}}$. Similar results hold for the eigenvalues and eigenvectors of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, as seen in (3.2.3).

All of the information that is available about $\boldsymbol{\beta}$ is carried by the reduced variables $\mathbf{L}^T\mathbf{Y}$ and $\mathbf{R}^T\mathbf{X}$, which can be seen as follows. Recall the singular value decomposition of $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\text{span}(\mathbf{U}) \subseteq \text{span}(\mathbf{L})$ and $\text{span}(\mathbf{V}) \subseteq \text{span}(\mathbf{R})$ by the definition of the $\mathbf{X}$- and $\mathbf{Y}$-envelopes. Hence, $\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^T = (\mathbf{L}\mathbf{A})\mathbf{D}(\mathbf{B}^T\mathbf{R}^T)$ for some semi-orthogonal matrices $\mathbf{A} \in \mathbb{R}^{d_Y\times d}$ and $\mathbf{B} \in \mathbb{R}^{d_X\times d}$. Then $\mathbf{L}_0^T\boldsymbol{\beta} = 0$, $\boldsymbol{\beta}\mathbf{R}_0 = 0$ and model (3.2.1) can be reduced to

$$\begin{cases} \mathbf{L}^T\mathbf{Y} = & \mathbf{L}^T\boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^T\{\mathbf{R}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})\} + \mathbf{L}^T\boldsymbol{\epsilon}, \\ \mathbf{L}_0^T\mathbf{Y} = & \mathbf{L}_0^T\boldsymbol{\mu}_{\mathbf{Y}} + \mathbf{L}_0^T\boldsymbol{\epsilon}, \end{cases} \tag{3.2.4}$$

where $\boldsymbol{\eta} = \mathbf{A}\mathbf{D}\mathbf{B}^T \in \mathbb{R}^{d_Y\times d_X}$ has rank d. The *simultaneous envelope model* then becomes

$$\begin{aligned} \mathbf{Y} &= \mathbf{L}\mathbf{L}^T\mathbf{Y} + \mathbf{L}_0\mathbf{L}_0^T\mathbf{Y} \\ &= \boldsymbol{\mu}_{\mathbf{Y}} + \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\epsilon}. && (3.2.5) \end{aligned}$$

with $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ given by (3.2.2) and (3.2.3).

Comparing to (3.2.1), we see that the regression coefficient matrix is now $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$, where $\boldsymbol{\eta}$ contains the coordinates of $\boldsymbol{\beta}$ relative to $\mathbf{L}$ and $\mathbf{R}$. This implies that the columns and rows of $\boldsymbol{\beta}$ vary only within the $\mathbf{Y}$-envelope and the $\mathbf{X}$-envelope. By letting $\mathbf{R} = \mathbf{I}_p$ or $\mathbf{L} = \mathbf{I}_r$, there will be reductions only in the column space or the row space of $\boldsymbol{\beta}$. These are the two special situations studied by Cook et al. (2010) and Cook et al. (2013).

If $\mathbf{R} = \mathbf{I}_p$, it follows from (3.2.3) and (3.2.4) that $\text{cov}(\mathbf{L}^T\mathbf{Y}, \mathbf{L}_0^T\mathbf{Y}|\mathbf{X}) = 0$ and $\mathbf{L}_0^T\mathbf{Y}|\mathbf{X} \sim \mathbf{L}_0^T\mathbf{Y}$, which motivated the construction of response envelopes (Cook, et al., 2010). If $\boldsymbol{\epsilon}$ is normally distributed, then this pair of conditions is equivalent to $\mathbf{L}_0^T\mathbf{Y} \perp\!\!\!\perp \mathbf{L}^T\mathbf{Y}|\mathbf{X}$ and $\mathbf{L}_0^T\mathbf{Y}|\mathbf{X} \sim \mathbf{L}_0^T\mathbf{Y}$. If $(\mathbf{X}, \mathbf{Y})$ is jointly normal then this pair of conditions is equivalent to

$\mathbf{L}_0^T\mathbf{Y} \perp\!\!\!\perp (\mathbf{L}^T\mathbf{Y}, \mathbf{X})$. We refer to $\mathbf{L}^T\mathbf{Y}$ and $\mathbf{L}_0^T\mathbf{Y}$ as the material and immaterial parts of the responses. This is because $\mathbf{L}_0^T\mathbf{Y}$ is neither affected by the predictors nor correlated with the complementary part of the responses $\mathbf{L}^T\mathbf{Y}$ and in this sense has no contribution to linear regression.

If $\mathbf{L} = \mathbf{I}_r$ then, from (3.2.2) and (3.2.5) , $\text{cov}(\mathbf{Y}, \mathbf{R}_0^T\mathbf{X}|\mathbf{R}^T\mathbf{X}) = 0$ and $\text{cov}(\mathbf{R}^T\mathbf{X}, \mathbf{R}_0^T\mathbf{X}) = 0$, which are the two conditions used by Cook, et al. (2013, Proposition 2.1) for predictor reduction. If $(\mathbf{X}, \mathbf{Y})$ has a joint normal distribution, then this pair of conditions is equivalent to $\mathbf{R}_0^T\mathbf{X} \perp\!\!\!\perp (\mathbf{R}^T\mathbf{X}, \mathbf{Y})$. We refer to $\mathbf{R}^T\mathbf{X}$ and $\mathbf{R}_0^T\mathbf{X}$ as the material and immaterial parts of the predictors. Similar to the response, this is because $\mathbf{R}_0^T\mathbf{X}$ is neither affected by the response nor correlated with the rest of the predictors.

The above conditions for $\mathbf{R}$ and $\mathbf{L}$ are not stated symmetrically because of the natural of regression. Cook et al. (2010) treated $\mathbf{X}$ as fixed since $\mathbf{S}_\mathbf{X}$ is an ancillary statistic for the $\mathbf{Y}$-envelope, while Cook et al. (2013) treated $\mathbf{X}$ as random because $\mathbf{S}_\mathbf{X}$ is not ancillary for $\mathbf{X}$ reduction. For simultaneous reductions, we assume that $\mathbf{X}$ and $\mathbf{Y}$ have a joint distribution throughout this article. The covariance decompositions (3.2.2) and (3.2.3) play a critical role in obtaining the above relationships, and they distinguish the envelope reductions from other methods for reducing the column and row dimensions of $\boldsymbol{\beta}$.

The previous relationships follow from the marginal response and predictor envelopes. The following lemma describes additional relationships between the material part in $\mathbf{X}$ (or $\mathbf{Y}$) and the immaterial part in $\mathbf{Y}$ (or $\mathbf{X}$) that come with the simultaneous envelopes.

Lemma 3.1. *Assume the simultaneous envelope model (3.2.5). Then* $\text{cov}(\mathbf{L}^T\mathbf{Y}, \mathbf{R}_0^T\mathbf{X}) = 0$ *and* $\text{cov}(\mathbf{R}^T\mathbf{X}, \mathbf{L}_0^T\mathbf{Y}) = 0$.

This lemma, which does not require normality of $\mathbf{X}$ or $\mathbf{Y}$, is implied by the previous discussion if $\mathbf{R} = \mathbf{I}_p$ or $\mathbf{L} = \mathbf{I}_r$. It shows a similarity between simultaneous envelope reduction and canonical correlation analysis: the selected components are uncorrelated with the rest of the components. Additional discussion of the connection between envelopes and canonical correlation is given in Section 3.2.3.

3.2.2 A visualized example of simultaneous envelope

The top two plots in Figure 3.1 illustrate the working mechanism of the simultaneous envelope reduction for a multivariate regression with two response $\mathbf{Y} = (Y_1, Y_2)^T$ and two predictors $\mathbf{X} = (X_1, X_2)^T$. For ease of illustration, we assume that $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$ with rank one regression coefficient matrix $\boldsymbol{\beta} = \mathbf{L}\mathbf{R}^T$ for some 2×1 matrices $\mathbf{R}$ and $\mathbf{L}$ such that $(\mathbf{R}, \mathbf{R}_0) \in \mathbb{R}^{2\times 2}$ and $(\mathbf{L}, \mathbf{L}_0) \in \mathbb{R}^{2\times 2}$ are orthogonal matrices. Then the plots demonstrate the set-up where $\mathbf{R}$ and $\mathbf{L}$ span the $\mathbf{X}$- and $\mathbf{Y}$-envelope.

In the first plot, the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is represented by the ellipses, whose axes are the directions of the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The shift from one contour to

another is captured by $\boldsymbol{\beta}\mathbf{X} = \mathbf{L}\mathbf{R}^T\mathbf{X}$, which is in the direction of $\mathbf{L}$ and has magnitude proportional to $\mathbf{R}^T\mathbf{X}$. From Proposition 1.1, the $\mathbf{Y}$-envelope is the sum of eigenspaces of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ that are not orthogonal to $\mathbf{L}$. In this plot, the eigenvector corresponds to the larger eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ is orthogonal to $\mathbf{L}$ and hence represents the immaterial information of the regression. By projecting the data onto $\mathbf{L}^T\mathbf{Y}$, we will eliminate the immaterial variation in the response. The response envelope reduction in this case is very efficient because $\mathbf{L}$ lies in the eigenspace corresponds to the small eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$.

The second plot represents the marginal distribution of $\mathbf{X}$. From Proposition 1.1, the reduction by $\mathbf{X}$-envelope is available when $\mathbf{R}$ is contained in a subset of the eigenspaces of $\boldsymbol{\Sigma}_{\mathbf{X}}$. In this plot, $\mathbf{R}$ happens to be the first eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}$, which means it spans the $\mathbf{X}$-envelope and $\mathbf{R}_0^T\mathbf{X}$ will be immaterial information of the regression. It should be pointed out that if we assign the Lebesgue measure to this 2-dimensional space, the probability of $\mathbf{R}$ being one of the two eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}$ will be zero. But statistically, this event could happen because it is equivalent to requiring (a) $\mathrm{cov}(\mathbf{Y}, \mathbf{R}_0^T\mathbf{X}|\mathbf{R}^T\mathbf{X}) = 0$ and (b) $\mathrm{cov}(\mathbf{R}^T\mathbf{X}, \mathbf{R}_0^T\mathbf{X}) = 0$, see discussion in Section 3.2.1. As represented by this plot, the predictor envelope reduction has great advantage over OLS when $\mathbf{R}$ corresponds to the larger eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{X}}$. This can be seen from the first plot, where the magnitude of $\mathbf{R}^T\mathbf{X}$ is proportional to the strength of the linear relationship.

For a toy data example, we use the meat data analyzed by Cook et al. (2013) for envelope predictor reduction in multivariate linear regression. This dataset consists of spectral measurements from infrared transmittance for 103 meat samples. There are three response variables: percentages of protein, fat and water. The sum of the three percentages is not one because of the other chemical content in the sample. We take Y_1 to be protein and Y_2 to be the sum of water and fat. We take two spectral measurements at 910nm and 960nm for illustration. The estimated $\mathbf{X}$-envelope and the $\mathbf{Y}$-envelope are both one-dimensional. The left plot was constructed by conditioning on the high, median and low values of $\widehat{\mathbf{R}}^T\mathbf{X}$, we can clearly see that the estimated envelope direction $\widehat{\mathbf{L}}$ matches the major axes of the contour of each sub-sample. So in this example the $\mathbf{Y}$-envelope reduces the dimension but not very much immaterial information. On the other hand, the right plot closely resembles the schematic representation shown in top-right. Consequently, the simultaneous envelope method offers a more precise estimate of $\boldsymbol{\beta}$ than the standard method. The standard errors of the OLS estimated coefficients in $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ are 1.2, 12.7, 12.8 and 49.5 times of that of the simultaneous envelope estimator.

3.2.3 Links to PCA, PLS, CCA and RRR

As stated previously, the eigenvalues of $\boldsymbol{\Omega}_0$ could be any subset of the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}}$. When some or all of the largest few eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}}$ come from $\boldsymbol{\Omega}_0$, the first few principal components of $\mathbf{X}$ will be from the immaterial part of $\mathbf{X}$, which is ineffective for

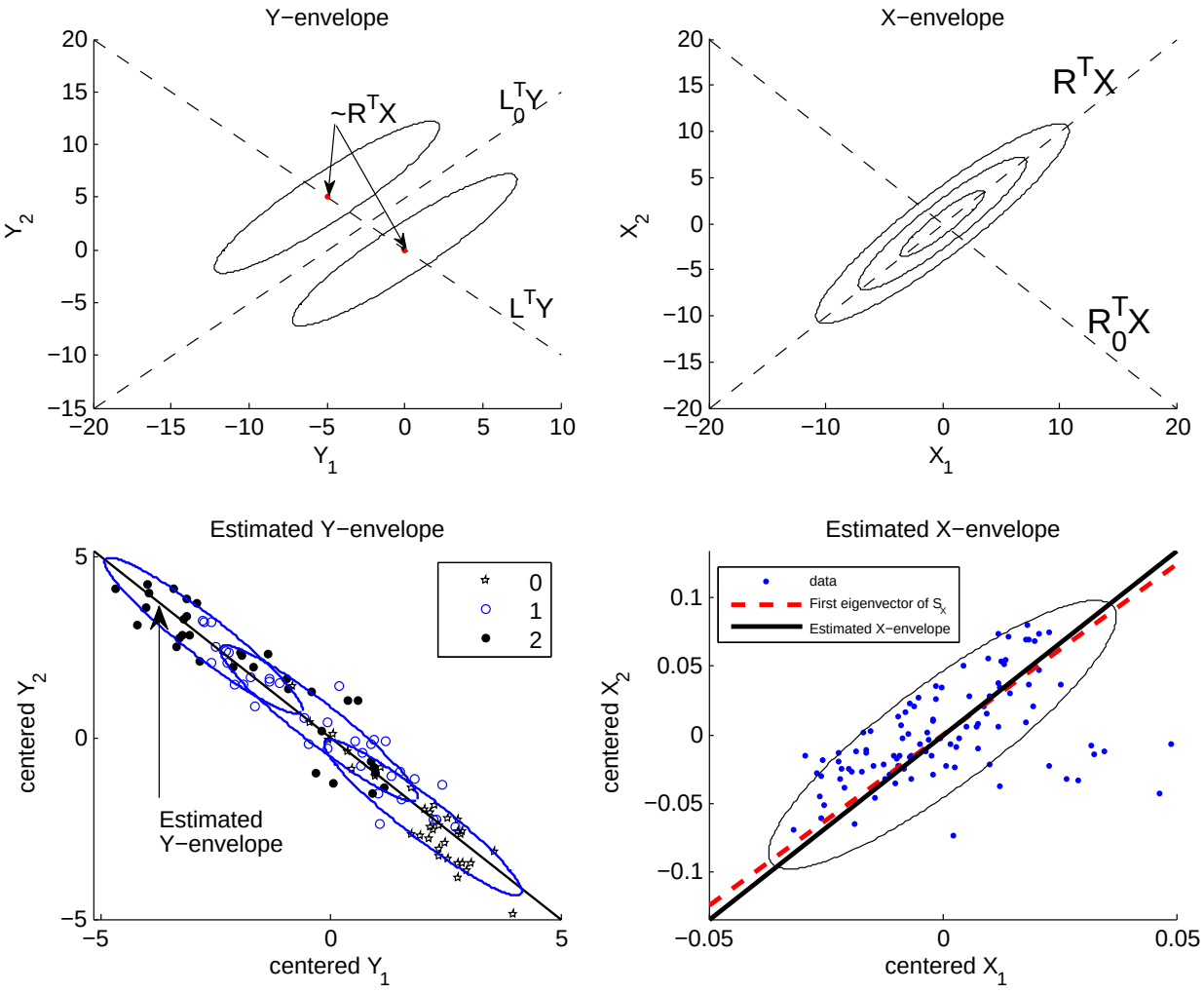


Figure 3.1: Working mechanism of simultaneous envelope reduction. Top left: Schematic representation of response envelope reduction; top right: schematic representation of envelope in the predictor space; bottom left: the meat data with estimated $\mathbf{Y}$-envelope, where the data points were marked differently according to their values in the predictor envelope $\widehat{\mathbf{R}}^T\mathbf{X}_i$; bottom right: the meat data with estimated $\mathbf{X}$-envelope and the first eigenvector of $\mathbf{S}_\mathbf{X}$. To help visualization, elliptical contours that cover 90% of the data points were added in the bottom two plots.

the regression as we can see from the above relationships. Principal components may be effective only if the larger eigenvalues of $\mathbf{\Sigma_X}$ all come from $\mathbf{\Omega}$; that is, from the material variation. Similar problems of principal component analysis arise for reducing $\mathbf{Y}$.

As mentioned in the Introduction, the partial least squares method reduces only the predictors and so it is comparable to simultaneous envelopes method when setting $\mathbf{L} = \mathbf{I}_r$. Cook, et al. (2013) showed that in this case the SIMPLS algorithm (de Jong, 1993) produces a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$, and that a likelihood-based estimator can do much better for prediction than the SIMPLS estimator.

Canonical correlation analysis is widely used for the purpose of simultaneously reducing the predictors and the responses. In the population, it finds canonical pairs of directions $\{\mathbf{a}_i, \mathbf{b}_i\}$, $i = 1, \ldots, d$, such that the correlations between $\mathbf{a}_i^T\mathbf{X}$ and $\mathbf{b}_i^T\mathbf{Y}$ are maximized. The maximization is over the constraints $\mathbf{a}_j^T\mathbf{\Sigma_X}\mathbf{a}_k = 0$, $\mathbf{a}_j^T\mathbf{\Sigma_{XY}}\mathbf{b}_k = 0$ and $\mathbf{b}_j^T\mathbf{\Sigma_Y}\mathbf{b}_k = 0$ for all $j \neq k$, and $\mathbf{a}_j^T\mathbf{\Sigma_X}\mathbf{a}_j = 1$ and $\mathbf{b}_j^T\mathbf{\Sigma_Y}\mathbf{b}_j = 1$ for all j. The solution is then $\{\mathbf{a}_i, \mathbf{b}_i\} = \{\mathbf{\Sigma_X}^{-1/2}\mathbf{e}_i, \mathbf{\Sigma_Y}^{-1/2}\mathbf{f}_i\}$, where $\{\mathbf{e}_i, \mathbf{f}_i\}$ is the i-th left-right eigenvector pair of the correlation matrix $\boldsymbol{\rho} = \mathbf{\Sigma_X}^{-1/2}\mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1/2}$.

Lemma 3.2. *Under the simultaneous envelope model (3.2.5), canonical correlation analysis can find at most d directions in the population, where $d = \mathrm{rank}(\boldsymbol{\beta}) \leq \min(d_X, d_Y)$. Moreover, the directions are contained in the simultaneous envelope as*

$$\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_d) \subseteq \mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T),\ \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_d) \subseteq \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta}). \tag{3.2.6}$$

Hence, canonical correlation analysis may miss some information about the regression by ignoring some material parts of $\mathbf{X}$ and/or $\mathbf{Y}$. For example, when r is small, it can find at most r linear combinations of $\mathbf{X}$, which can be insufficient for regression. Moreover, the most correlated pairs are not, in general, the most predictable pairs for regression. Canonical correlations are often used for data visualization instead of regression, so it may be expected that they can fail in prediction. In the simulation studies of Section 3.6.1 and Section 3.6.2 in the Supplement, we found that the prediction performances based on canonical correlations varied widely for different covariance structures.

The maximum likelihood estimator of RRR is obtained as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RR}} = \arg\min_{\mathrm{rank}(\boldsymbol{\beta})=d} \{\sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i)^T \mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1} (\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i)\}, \tag{3.2.7}$$

which is the same as the regression coefficient matrix obtained by using the canonical variables (see Reinsel and Velu 1998; Section 2.4.2). Therefore, in the purpose of estimation and prediction, the maximum likelihood RRR estimator is equivalent to the estimator based on CCA estimator. RRR can also be applied with identity inner product in (3.2.7) instead of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. And for any minimization criteria of RRR, the estimators always have the form $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}$ where $\widehat{\mathbf{A}} \in \mathbb{R}^{r\times d}$ and $\widehat{\mathbf{B}} \in \mathbb{R}^{d\times p}$ are both $\sqrt{n}$-consistent estimators

for their population counterparts $\mathbf{A}$ and $\mathbf{B}$. Similar to Lemma 3.2, we have the following relations by definition,

$$\text{span}(\mathbf{A}) = \text{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}),\ \text{span}(\mathbf{B}^T) = \text{span}(\boldsymbol{\beta}^T) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T). \tag{3.2.8}$$

Therefore, RRR has the same drawback as CCA that it may loss some material information in regression. Simulation studies in Section 3.6 includes RRR estimator based on identity inner product.

3.2.4 Potential gain

To gain intuition about the potential advantages of simultaneous envelopes, we next consider the case where $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ are known. Estimation of these envelopes in practice will mitigate the findings in this section, but we have nevertheless found them to be useful qualitative indicators of the benefits of simultaneous reduction.

The envelope estimator for $\boldsymbol{\beta}$ with known semi-orthogonal basis matrices $\mathbf{R}$ and $\mathbf{L}$, denoted by $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$, can be written as

$$\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}} = \mathbf{L}\widehat{\boldsymbol{\eta}}_{\mathbf{R},\mathbf{L}}\mathbf{R}^T = \mathbf{L}\mathbf{L}^T\mathbf{S}_{\mathbf{YX}}\mathbf{R}(\mathbf{R}^T\mathbf{S}_{\mathbf{X}}\mathbf{R})^{-1}\mathbf{R}^T = \mathbf{P}_{\mathbf{L}}\widehat{\boldsymbol{\beta}}_{\text{OLS}}\mathbf{P}^T_{\mathbf{R}(\mathbf{S}_{\mathbf{X}})}, \tag{3.2.9}$$

where $\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}$ is the ordinary least squares estimator. Clearly, the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ uses only the material variation in $\mathbf{Y}$ and $\mathbf{X}$. It can be obtained by projecting $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ to the reduced predictor space and the reduced response space, and so does not depend on the particular bases $\mathbf{R}$ and $\mathbf{L}$ selected. The estimator $\widehat{\boldsymbol{\eta}}_{\mathbf{R},\mathbf{L}}$ is the ordinary least squares estimator of the coefficient matrix for the regression of $\mathbf{L}^T\mathbf{Y}$ on $\mathbf{R}^T\mathbf{X}$.

The next proposition shows how the variance of the ordinary least squares estimator with normal predictors can be potentially reduced by using simultaneous envelopes. Let $f_p = n - p - 2$ and $f_x = n - d_X - 2$.

Proposition 3.1. *Assume that* $\mathbf{X} \sim N_p(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, $n > p + 2$ *and that semi-orthogonal basis matrices* $\mathbf{R}$, $\mathbf{L}$ *for the left and right envelopes are known. Then* $\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) = f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ *and*

$$\begin{aligned}
\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})) &= f_x^{-1}\left(\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T\right) \otimes \left(\mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T\right) \\
&= f_p f_x^{-1}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) - f_x^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T \\
&\quad - f_x^{-1}\mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T - f_x^{-1}\mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T,
\end{aligned}$$

where $\boldsymbol{\Omega} = \mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}$, $\boldsymbol{\Omega}_0 = \mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}_0$, $\boldsymbol{\Phi}_0 = \mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}_0$ *and* $\boldsymbol{\Phi} = \mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}$.

This proposition shows that the variation in $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ can be seen in two parts: the first part is the variation in $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ times a constant $f_p f_x^{-1} \leq 1$, and the second consists of terms that reduce this value depending on the variances associated with the immaterial information $\mathbf{R}_0^T\mathbf{X}$ and $\mathbf{L}_0^T\mathbf{Y}$.

If $\mathbf{L} = \mathbf{I}_r$ then $\mathbf{\Phi} = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and we get the multivariate version of Proposition 2.3 by Cook, et al. (2013) for univariate response regression:

$$\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R}})) = f_p f_x^{-1}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) - f_x^{-1}\mathbf{R}_0\mathbf{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}. \tag{3.2.10}$$

When p is close to n and the $\mathbf{X}$-envelope dimension d_X is small, the constant $f_p f_x^{-1}$ could be small and the gain by $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ over $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ could be substantial. If there is substantial colinearity in the predictors, so $\mathbf{\Sigma}_{\mathbf{X}}$ has some small eigenvalues, and if the corresponding eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ fall in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, then the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ could be reduced considerably since $\mathbf{\Omega}_0^{-1}$ will be large. It is widely known that colinearity in $\mathbf{X}$ can increase the variance of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. However, when the eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ corresponding to these small eigenvalues lie in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ is not affected by colinearity.

Similarly, if $\mathbf{R} = \mathbf{I}_p$ then $\mathbf{\Omega} = \mathbf{\Sigma}_{\mathbf{X}}$ and we get the following new expression for $\mathbf{Y}$ reduction:

$$\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L}})) = f_p f_x^{-1}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) - f_x^{-1}\mathbf{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T. \tag{3.2.11}$$

If the eigenvectors with larger eigenvalues of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ lie in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$, then the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{L}}$ may be reduced considerably since then $\mathbf{\Phi}_0$ will be large.

More importantly, the last term of the expansion in Proposition 3.1 represents a synergy between the $\mathbf{X}$ and $\mathbf{Y}$ reductions that is not present in the marginal reductions. If the eigenvectors of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ with large eigenvalues lie in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ or if the eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ with small eigenvalues lie in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$, then the variance reductions in either (3.2.11) or (3.2.10) could be insignificant. However, the synergy in simultaneous $\mathbf{X}$ and $\mathbf{Y}$ reductions may still reduce the variance substantially because one of factors in the Kronecker product $f_x^{-1}\mathbf{R}_0\mathbf{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T$ could be still be large.

Let $\mathbf{x}_N$ denote a new observation from $N(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}})$, let $\mathbf{z}_N = \mathbf{x}_N - \boldsymbol{\mu}_{\mathbf{X}}$ be held fixed and consider $\text{var}(\widehat{\boldsymbol{\beta}} \cdot \mathbf{z}_N) = \text{var}\left((\mathbf{z}_N^T \otimes \mathbf{I})\text{vec}(\widehat{\boldsymbol{\beta}})\right)$, which is the variance of a fitted vector for $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$. It is straightforward from Proposition 3.1 that,

$$\begin{aligned} f_p\text{var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}} \cdot \mathbf{z}_N) &= f_x\text{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}} \cdot \mathbf{z}_N) + \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T\left(\mathbf{z}_N^T\mathbf{R}\mathbf{\Omega}^{-1}\mathbf{R}^T\mathbf{z}_N\right) \\ &\quad + \mathbf{L}\mathbf{\Phi}\mathbf{L}^T(\mathbf{z}_N^T\mathbf{R}_0\mathbf{\Omega}_0^{-1}\mathbf{R}_0^T\mathbf{z}_N) + \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T\left(\mathbf{z}_N^T\mathbf{R}_0\mathbf{\Omega}_0^{-1}\mathbf{R}_0^T\mathbf{z}_N\right). \end{aligned}$$

We see from the above equality that only the part of $\mathbf{z}_N$ that lies in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ will contribute to the variance in prediction from $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$, while the prediction variance from $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ depends on the whole of $\mathbf{z}_N$. If, in an extreme case, $\mathbf{z}_N \in \mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, then $\text{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}} \cdot \mathbf{z}_N) = 0$.

3.3 Estimating envelopes

Let $\mathbf{C} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ denote the concatenated random vectors, which has mean $\boldsymbol{\mu}_{\mathbf{C}}$ and covariance $\mathbf{\Sigma}_{\mathbf{C}}$, and let $\mathbf{S}_{\mathbf{C}}$ be the sample covariance matrix of $\mathbf{C}$. In order to estimate

the parameters of the simultaneous envelope model (3.2.5), we introduce and probe a likelihood-based objective function that includes the objective functions in Cook, et al. (2010) and Cook, et al. (2013) as special cases. Variations on the objective function and their corresponding algorithms are also studied. We first give coordinate dependent and coordinate independent representations of $\mathbf{\Sigma_C}$ in Section 3.3.1 to facilitate estimation.

3.3.1 Structure of the covariances

Since we already have the structure of $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_{Y|X}}$ in (3.2.2) and (3.2.3), and $\mathbf{\Sigma_{XY}} = \mathbf{\Sigma_X}\boldsymbol{\beta}^T = \mathbf{R\Omega\eta}^T\mathbf{L}^T$, we need only the following expression for $\mathbf{\Sigma_Y}$ to complete the necessary ingredients of $\mathbf{\Sigma_C}$,

$$\begin{aligned}
\mathbf{\Sigma_Y} &= \mathbf{\Sigma_{Y|X}} + \mathbf{\Sigma_{XY}^T}\mathbf{\Sigma_X^{-1}}\mathbf{\Sigma_{XY}} \\
&= \mathbf{L\Phi L}^T + \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T + \mathbf{L\eta\Omega R}^T\left(\mathbf{R\Omega R}^T + \mathbf{R}_0\mathbf{\Omega}_0\mathbf{R}_0^T\right)^{-1}\mathbf{R\Omega\eta}^T\mathbf{L}^T \\
&= \mathbf{L}(\mathbf{\Phi} + \boldsymbol{\eta\Omega\eta}^T)\mathbf{L}^T + \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T.
\end{aligned}$$

Then we get the coordinate representation of the covariance $\mathbf{\Sigma_C}$ as

$$\begin{aligned}
\mathbf{\Sigma_C} &= \begin{pmatrix} \mathbf{\Sigma_X} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{XY}^T} & \mathbf{\Sigma_Y} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{R\Omega R}^T + \mathbf{R}_0\mathbf{\Omega}_0\mathbf{R}_0^T & \mathbf{R\Omega\eta}^T\mathbf{L}^T \\ \mathbf{L\eta\Omega R}^T & \mathbf{L}(\mathbf{\Phi} + \boldsymbol{\eta\Omega\eta}^T)\mathbf{L}^T + \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T \end{pmatrix}. \qquad (3.3.1)
\end{aligned}$$

By noticing that $\mathbf{\Sigma_X} = \mathbf{P_R\Sigma_X P_R} + \mathbf{Q_R\Sigma_X Q_R}$, $\mathbf{\Sigma_{XY}} = \mathbf{P_R\Sigma_{XY}P_L}$ and $\mathbf{\Sigma_Y} = \mathbf{P_L\Sigma_Y P_L} + \mathbf{Q_L\Sigma_Y Q_L}$ from the above expression, we can further obtain the coordinate independent representation

$$\begin{aligned}
\mathbf{\Sigma_C} &= \mathbf{P_{R\oplus L}\Sigma_C P_{R\oplus L}} + \mathbf{Q_{R\oplus L}\Sigma_C Q_{R\oplus L}}, \\
&= \mathbf{P_{R\oplus L}\Sigma_C P_{R\oplus L}} + \mathbf{Q_{R\oplus L}\Sigma_D Q_{R\oplus L}}, \qquad (3.3.2)
\end{aligned}$$

where $\mathbf{\Sigma_D} \equiv \mathbf{\Sigma_X} \oplus \mathbf{\Sigma_Y}$ and $\mathbf{P_{R\oplus L}} = \mathbf{P_R} \oplus \mathbf{P_L}$ is the projection onto the direct sum of two envelopes, $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) \oplus \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$.

So far we have considered $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$ and $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta})$ as separate subspaces. Motivated by (3.3.2), the next lemma states that the direct sum of two arbitrary envelopes is itself an envelope. Let $\mathbf{M}_1 \in \mathbb{S}^{p_1\times p_1}$, $\mathbf{M}_2 \in \mathbb{S}^{p_2\times p_2}$, and let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces of $\text{span}(\mathbf{M}_1)$ and $\text{span}(\mathbf{M}_2)$, which is required by Definition 1.2. Then

Lemma 3.3. $\mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1) \oplus \mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2) = \mathcal{E}_{\mathbf{M}_1\oplus\mathbf{M}_2}(\mathcal{S}_1 \oplus \mathcal{S}_2)$.

From this lemma, we have $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) \oplus \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta}) = \mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) \oplus \mathcal{E}_{\mathbf{\Sigma_Y}}(\boldsymbol{\beta}) = \mathcal{E}_{\mathbf{\Sigma_X}\oplus\mathbf{\Sigma_Y}}(\boldsymbol{\beta}^T \oplus \boldsymbol{\beta}) = \mathcal{E}_{\mathbf{\Sigma_D}}(\boldsymbol{\beta}^T \oplus \boldsymbol{\beta})$. We call this the simultaneous envelope for $\boldsymbol{\beta}$.

3.3.2 The estimation criterion and resulting estimators

Assuming multivariate normality for $\mathbf{C}$, the negative log-likelihood minimized over $\boldsymbol{\mu}_{\mathbf{C}}$ leads to the following objective function for estimation of $\boldsymbol{\Sigma}_{\mathbf{C}}$,

$$\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}}) = \log|\boldsymbol{\Sigma}_{\mathbf{C}}| + \mathrm{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}). \tag{3.3.3}$$

We use this as a multi-purpose objective function, which gives the maximum likelihood estimator of $\boldsymbol{\beta}$ under normality of $\mathbf{C}$ and a $\sqrt{n}$-consistent simultaneous envelope estimator of $\boldsymbol{\beta}$ when $\mathbf{C}$ has finite fourth moments. We also use $\mathrm{F}(\cdot)$ as a generic objective function whose definition changes and is implied by its own argument. This should cause no confusion since $\mathrm{F}(\cdot)$ will always be written with its arguments.

Substituting the coordinate form for $\boldsymbol{\Sigma}_{\mathbf{C}}$ from (3.3.1) into $\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$ leads to an objective function that can be minimized explicitly over $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$ and $\boldsymbol{\eta}$ with $\mathbf{L}$ and $\mathbf{R}$ held fixed. The resulting partially minimized objective function for simultaneous envelopes can be expressed as follows.

$$\mathrm{F}(\mathbf{R}\oplus\mathbf{L}) = \log|(\mathbf{R}^T\oplus\mathbf{L}^T)\mathbf{S}_{\mathbf{C}}(\mathbf{R}\oplus\mathbf{L})| + \log|(\mathbf{R}^T\oplus\mathbf{L}^T)\mathbf{S}_{\mathbf{D}}^{-1}(\mathbf{R}\oplus\mathbf{L})|, \tag{3.3.4}$$

where $\mathbf{S}_{\mathbf{D}} = \mathbf{S}_{\mathbf{X}}\oplus\mathbf{S}_{\mathbf{Y}}$ is the sample version of $\boldsymbol{\Sigma}_{\mathbf{D}}$. Let $|\mathbf{A}|_0$ denote the product of all nonzero eigenvalues of a matrix $\mathbf{A}$. Then we have the following coordinate independent representation of (3.3.4),

$$\mathrm{F}(\mathbf{P}_{\mathbf{R}\oplus\mathbf{L}}) = \log|\mathbf{P}_{\mathbf{R}\oplus\mathbf{L}}\mathbf{S}_{\mathbf{C}}\mathbf{P}_{\mathbf{R}\oplus\mathbf{L}}|_0 + \log|\mathbf{P}_{\mathbf{R}\oplus\mathbf{L}}\mathbf{S}_{\mathbf{D}}^{-1}\mathbf{P}_{\mathbf{R}\oplus\mathbf{L}}|_0. \tag{3.3.5}$$

Moreover, if we substitute the coordinate free representation of $\boldsymbol{\Sigma}_{\mathbf{C}}$ from (3.3.2) into $\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$, we immediately get (3.3.5).

Objective function (3.3.4) is invariant under orthogonal transformations: for any orthogonal $(d_X+d_Y)\times(d_X+d_Y)$ matrix $\mathbf{O}$, $\mathrm{F}(\mathbf{R}\oplus\mathbf{L}) = \mathrm{F}((\mathbf{R}\oplus\mathbf{L})\mathbf{O})$. The same result holds if we replace $\mathbf{S}_{\mathbf{C}}$ and $\mathbf{S}_{\mathbf{D}}$ with their population counterparts. Minimization of $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ is thus a Grassmann optimization problem with special direct sum structure. Consequently, neither $\mathbf{R}$ nor $\mathbf{L}$ is identified. However, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ are identified, and these are all that is needed to estimate $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. While the estimators of $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_0$ depend on the particular bases chosen to minimize $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$, the estimators of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ are independent of the choice since the estimated projection matrices $\mathbf{P}_{\widehat{\mathbf{R}}}$ and $\mathbf{P}_{\widehat{\mathbf{L}}}$ do not depend on the basis. In short, any values $\widehat{\mathbf{R}}$ of $\mathbf{R}$ and $\widehat{\mathbf{L}}$ of $\mathbf{L}$ that minimize $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ are allowed. Estimators of $\mathbf{R}_0$ and $\mathbf{L}_0$ are then any semi-orthogonal matrices $\widehat{\mathbf{R}}_0$ and $\widehat{\mathbf{L}}_0$ so that $(\widehat{\mathbf{R}},\widehat{\mathbf{R}}_0)$ and $(\widehat{\mathbf{L}},\widehat{\mathbf{L}}_0)$ are orthogonal matrices.

The next lemma summarizes the estimators that result from this process.

Lemma 3.4. *Let* $\widehat{\mathbf{R}}\oplus\widehat{\mathbf{L}} = \arg\min \mathrm{F}(\mathbf{R}\oplus\mathbf{L})$, *where* $\mathbf{R}\in\mathbb{R}^{p\times d_X}$ *and* $\mathbf{L}\in\mathbb{R}^{r\times d_Y}$ *are semi-orthogonal basis matrices. Then the estimators of the remaining parameters are*

$\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{R}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{R}}$, $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\mathbf{R}}_0^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{R}}_0$, $\widehat{\boldsymbol{\eta}} = (\widehat{\mathbf{L}}^T\mathbf{S}_{\mathbf{YX}}\widehat{\mathbf{R}})(\widehat{\mathbf{R}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{R}})^{-1}$, $\widehat{\boldsymbol{\Phi}}_0 = \widehat{\mathbf{L}}_0^T\mathbf{S}_{\mathbf{Y}}\widehat{\mathbf{L}}_0$ *and*

$$\widehat{\boldsymbol{\Phi}} = \widehat{\mathbf{L}}^T\left(\mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{XY}}^T\widehat{\mathbf{R}}\left(\widehat{\mathbf{R}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{R}}\right)^{-1}\widehat{\mathbf{R}}^T\mathbf{S}_{\mathbf{XY}}\right)\widehat{\mathbf{L}}.$$

The simultaneous envelope estimator for $\boldsymbol{\beta}$ is then

$$\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{L}}\widehat{\boldsymbol{\eta}}\widehat{\mathbf{R}}^T = \mathbf{P}_{\widehat{\mathbf{L}}}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\mathbf{P}^T_{\widehat{\mathbf{R}}(\mathbf{S}_{\mathbf{X}})}. \tag{3.3.6}$$

The simultaneous envelope estimator for $\boldsymbol{\beta}$ in (3.3.6) coincides with the plug-in envelope estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathbf{R}},\widehat{\mathbf{L}}}$ obtained by regarding estimated $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{L}}$ as the known values of $\mathbf{R}$ and $\mathbf{L}$. We next turn to methods for minimizing (3.3.4).

3.3.3 Alternating algorithm

If we fix $\mathbf{R}$ as an arbitrary orthogonal basis, then the objective function $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ in (3.3.4) can be re-expressed as an objective function in $\mathbf{L}$:

$$\mathrm{F}(\mathbf{L}|\mathbf{R}) = \log|\mathbf{L}^T\mathbf{S}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}\mathbf{L}| + \log|\mathbf{L}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{L}|. \tag{3.3.7}$$

Similarly, if we fix $\mathbf{L}$, the objective function $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ reduces to the conditional objective function

$$\mathrm{F}(\mathbf{R}|\mathbf{L}) = \log|\mathbf{R}^T\mathbf{S}_{\mathbf{X}|\mathbf{L}^T\mathbf{Y}}\mathbf{R}| + \log|\mathbf{R}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{R}|. \tag{3.3.8}$$

We use the following alternating algorithm based on (3.3.7) and (3.3.8) to obtain a minimizer of the objective function $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ in (3.3.4).

1. Initialization. Set the starting value $\mathbf{R}^{(0)}$ and get $\mathbf{L}^{(0)} = \arg\min_{\mathbf{L}}\mathrm{F}(\mathbf{L}|\mathbf{R}^{(0)})$.

2. Alternating. For the k-th stage, obtain $\mathbf{R}^{(k)} = \arg\min_{\mathbf{R}}\mathrm{F}(\mathbf{R}|\mathbf{L} = \mathbf{L}^{(k-1)})$ and $\mathbf{L}^{(k)} = \arg\min_{\mathbf{L}}\mathrm{F}(\mathbf{L}|\mathbf{R} = \mathbf{R}^{(k)})$.

3. Convergence criterion. Evaluate $\left\{\mathrm{F}(\mathbf{R}^{(k-1)}\oplus\mathbf{L}^{(k-1)}) - \mathrm{F}(\mathbf{R}^{(k)}\oplus\mathbf{L}^{(k)})\right\}$ and return to the alternating step if it is bigger than a tolerance; otherwise, stop the iteration and use $\mathbf{R}^{(k)}\oplus\mathbf{L}^{(k)}$ as the final estimator.

In the initialization step of the algorithm, we could also set $\mathbf{L}^{(0)}$ to be some initial value and get $\mathbf{R}^{(0)} = \arg\min_{\mathbf{R}}\mathrm{F}(\mathbf{R}|\mathbf{L}^{(0)})$. Then interchanging the roles of $\mathbf{R}$ and $\mathbf{L}$ in the alternating step will give us another algorithm, which has the same performance as the alternating algorithm outlined above. Comparing to the objective function used by Cook et al. (2010), we see that $\mathrm{F}(\mathbf{L}|\mathbf{R})$ is the objective function for estimating the $\mathbf{Y}$-envelope in the regression of $\mathbf{Y}$ on the reduced predictors $\mathbf{R}^T\mathbf{X}$ for fixed $\mathbf{R}$. Similarly from the objective function used by Cook et al. (2013), we notice that $\mathrm{F}(\mathbf{R}|\mathbf{L})$ is the objective function for estimating the $\mathbf{X}$-envelope in the regression of the reduced responses $\mathbf{L}^T\mathbf{Y}$ on the predictors $\mathbf{X}$ for fixed $\mathbf{L}$.

In our experience, as long as we use good initial values, the alternating algorithm, which monotonically decreases $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$, will converge after only a few cycles, typically less than four. Root-n consistent starting values are particularly important to mitigate potential problems caused by multiple local minima and to ensure efficient estimation. For instance, under joint normality of $\mathbf{X}$ and $\mathbf{Y}$, one Newton-Raphson iteration from any $\sqrt{n}$-consistent estimator will be asymptotically as efficient as the maximum likelihood estimator, even if local minima are present (Small et al., 2000; Lehmann and Casella, 1998, Theorem 4.3).

The 1D algorithm in Chapter 2, Algorithm 1, can be used to get fast $\sqrt{n}$-consistent starting values for the alternating algorithm by separately estimating the $\mathbf{X}$-envelope and the $\mathbf{Y}$-envelope bases. Since the 1D algorithm turns the optimizations over d_X and d_Y-dimensional manifolds to d_X and d_Y sequential optimizations over one-dimensional manifolds, the computation complexity is reduced drastically. For the simulation examples we studied, the computation costs of the 1D algorithm are from tens to hundreds times less than the computation costs of using d_X and d_Y-dimensional Grassmann manifold optimizations.

Perhaps more importantly, we have found $\widehat{\mathbf{R}}_{1\mathrm{D}}$ and $\widehat{\mathbf{L}}_{1\mathrm{D}}$, from optimizing $\mathrm{F}(\mathbf{R}|\mathbf{L}=\mathbf{I}_r)$ and $\mathrm{F}(\mathbf{L}|\mathbf{R}=\mathbf{I}_p)$ separately using the 1D algorithm, to be practically as efficient as the final estimators obtained by alternating because the alternating algorithm nearly always converges after only a few iterations and produces only small changes. To emphasize the utility of the 1D estimators, we use them in the simulations and real data examples that follow in later sections. In Section 3.6.4, we use a simulation example to demonstrate that estimators from the 1D algorithm may have the same behavior as maximum likelihood estimators.

3.4 Asymptotic properties

The parameters involved in the coordinate representation of the simultaneous envelope model are $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$, $\mathbf{R}$ and $\mathbf{L}$. In Sections 3.4.1 and 3.4.2, we focus on the asymptotic properties of the estimable functions $\boldsymbol{\beta}=\mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$, $\boldsymbol{\Sigma}_{\mathbf{X}}=\mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T+\mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T$, and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}=\mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T+\mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T$. Specifically, we study the asymptotic covariances of the following parameters $\boldsymbol{\phi}$ and estimable functions $\boldsymbol{h}$.

$$\boldsymbol{\phi}=\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\\ \boldsymbol{\phi}_3\\ \boldsymbol{\phi}_4\\ \boldsymbol{\phi}_5\\ \boldsymbol{\phi}_6\\ \boldsymbol{\phi}_7\end{pmatrix}\equiv\begin{pmatrix}\mathrm{vec}(\boldsymbol{\eta})\\ \mathrm{vec}(\mathbf{R})\\ \mathrm{vec}(\mathbf{L})\\ \mathrm{vech}(\boldsymbol{\Omega})\\ \mathrm{vech}(\boldsymbol{\Omega}_0)\\ \mathrm{vech}(\boldsymbol{\Phi})\\ \mathrm{vech}(\boldsymbol{\Phi}_0)\end{pmatrix},\ \boldsymbol{h}=\begin{pmatrix}\boldsymbol{h}_1(\boldsymbol{\phi})\\ \boldsymbol{h}_2(\boldsymbol{\phi})\\ \boldsymbol{h}_3(\boldsymbol{\phi})\end{pmatrix}\equiv\begin{pmatrix}\mathrm{vec}(\boldsymbol{\beta})\\ \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})\\ \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})\end{pmatrix}.$$

Since $\boldsymbol{h} \in \mathbb{R}^{\frac{1}{2}(p+r)(p+r+1)}$ and $\boldsymbol{\phi} \in \mathbb{R}^{\frac{1}{2}p(p+1)+\frac{1}{2}r(r+1)+d_X d_Y}$, the simultaneous envelope model reduces the total number of parameters by $\frac{1}{2}(p+r)(p+r+1) - \frac{1}{2}p(p+1) - \frac{1}{2}r(r+1) - d_X d_Y = pr - d_X d_Y$.

3.4.1 Without the normality assumption

Let $\widehat{\boldsymbol{h}}_{\text{full}} = \left(\text{vec}^T(\widehat{\boldsymbol{\beta}}_{\text{OLS}}), \text{vech}^T(\mathbf{S}_{\mathbf{X}}), \text{vech}^T(\mathbf{S}_{\mathbf{Y}|\mathbf{X}})\right)^T$ denote the OLS estimator of $\boldsymbol{h}$ under the standard model (3.2.1), and let $\widehat{\boldsymbol{h}}$ denote the estimator from Lemma 3.4 under the simultaneous envelope model (3.2.5). The true values of $\boldsymbol{h}$ and $\boldsymbol{\phi}$ are denoted as $\boldsymbol{h}_0$ and $\boldsymbol{\phi}_0$. Define $\boldsymbol{\Delta} \equiv \partial \boldsymbol{h}(\boldsymbol{\phi})/\partial \boldsymbol{\phi}$ to be the gradient matrix, whose explicit form is given in the Supplement, Section 3.9.7. We use "avar" to denote an asymptotic covariance matrix: $\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) = \mathbf{A}$ is equivalent to $\sqrt{n}(\widehat{\boldsymbol{h}} - \boldsymbol{h}_0) \to N(0, \mathbf{A})$ in distribution. We use the expansion matrix (Henderson and Searle, 1979) $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ to relate the vec operation and vech operation: for any symmetric matrix $\mathbf{M} \in \mathbb{S}^{p\times p}$, $\text{vec}(\mathbf{M}) = \mathbf{E}_p \text{vech}(\mathbf{M})$. Then $\sqrt{n}(\widehat{\boldsymbol{h}}_{\text{full}} - \boldsymbol{h}_0) \to N(0, \boldsymbol{\Gamma})$, for some positive definite covariance matrix $\boldsymbol{\Gamma}$. Since there is a one-to-one relationship between h and $\boldsymbol{\Sigma}_{\mathbf{C}}$, we treat the objective function $\text{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$ in (3.3.3) as a function of $\boldsymbol{h}$ and $\widehat{\boldsymbol{h}}_{\text{full}}$ and write it as $\text{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})$. Let $\mathbf{J}_{\boldsymbol{h}} = 1/2 \times \partial^2 \text{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})/\partial \boldsymbol{h} \partial \boldsymbol{h}^T$ evaluated at $\widehat{\boldsymbol{h}}_{\text{full}} = \boldsymbol{h} = \boldsymbol{h}_0$, which is the Fisher information matrix for $\boldsymbol{h}$ when $\mathbf{C}$ is normal,

$$\mathbf{J}_{\boldsymbol{h}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} & 0 & 0 \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p & 0 \\ 0 & 0 & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1})\mathbf{E}_r \end{pmatrix}. \tag{3.4.1}$$

The following proposition formally states the asymptotic distribution of $\widehat{\boldsymbol{h}}$.

Proposition 3.2. *Assume that the data $(\mathbf{x}_i, \mathbf{y}_i)$ are i.i.d. from a joint distribution with finite fourth moments. Then $\sqrt{n}(\text{vec}(\widehat{\boldsymbol{h}}) - \text{vec}(\boldsymbol{h}_0))$ converges in distribution to a normal random variable with mean $\mathbf{0}$ and covariance matrix*

$$\mathbf{W} = \boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T \mathbf{J}_{\boldsymbol{h}} \boldsymbol{\Delta}\right)^{\dagger} \boldsymbol{\Delta}^T \mathbf{J}_{\boldsymbol{h}} \boldsymbol{\Gamma} \mathbf{J}_{\boldsymbol{h}} \boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T \mathbf{J}_{\boldsymbol{h}} \boldsymbol{\Delta}\right)^{\dagger} \boldsymbol{\Delta}^T,$$

where $\dagger$ indicates the Moore-Penrose inverse. In particular, $\sqrt{n}(\text{vec}(\widehat{\boldsymbol{\beta}}) - \text{vec}(\boldsymbol{\beta}))$ converges in distribution to a normal random variable with mean $\mathbf{0}$ and covariance $\mathbf{W}_{11}$, the upper-left $pr \times pr$ block of $\mathbf{W}$. Moreover, $\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{full}}) \geq \text{avar}(\sqrt{n}\widehat{\boldsymbol{h}})$ if $\text{span}(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta})$ is a reducing subspace of $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}$.

The $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\beta}}$ is essentially because $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{XY}}$ are $\sqrt{n}$-consistent and because of the properties of $\text{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})$. The asymptotic covariance matrix $\mathbf{W}_{11}$ can be computed straightforwardly, but its accuracy for any fixed sample size may depend on the distribution of $\mathbf{C}$. Fortunately, bootstrap methods can provide a good approximation of $\mathbf{W}_{11}$, as discussed in Section 3.4.3.

3.4.2 Under the normality assumption

The asymptotic covariance $\mathbf{W}$ simplifies when $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$ because then objective function $\mathrm{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})$ agrees with the negative log-likelihood function and $\boldsymbol{\Gamma} = \mathbf{J}_{\boldsymbol{h}}^{-1} = \text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{full}})$. And $\{\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{full}}) - \text{avar}(\sqrt{n}\widehat{\boldsymbol{h}})\}$ will be positive semi-definite because $\text{span}(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta})$ is always a reducing subspace of $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2} = \mathbf{I}$.

Proposition 3.3. *Assume that* $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$. *Then* $\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T$. *Moreover,* $\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) \leq \text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{full}})$,

$$\mathbf{J}_{\boldsymbol{h}}^{-1} - \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T = \mathbf{J}_{\boldsymbol{h}}^{-\frac{1}{2}}\mathbf{Q}_{\mathbf{J}_{\boldsymbol{h}}^{\frac{1}{2}}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{-\frac{1}{2}} \geq 0.$$

In particular, $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})) \leq \text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}))$.

This proposition is obtained by direct computation, and is consistent with Cook et al. (2010) and Cook et al. (2013). Also, because the estimator $\widehat{\boldsymbol{\beta}}$ is the MLE under normality, its asymptotic variance is no larger than that of $\mathbf{X}$-envelope estimator or $\mathbf{Y}$-envelope estimator. Explicit expressions can be found in the Supplement, Section 3.9.7. To further interpret this result, we next consider the asymptotic variance of $\text{vec}(\widehat{\boldsymbol{\beta}})$, which can lead to the asymptotic variances for predictions and fitted values.

Proposition 3.4. *Assume that* $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$. *Then*

$$\begin{aligned}\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})) &= \text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})) + \text{avar}\left(\sqrt{n}\text{vec}(\mathbf{Q}_{\mathbf{L}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}})\right) \\ &+ \text{avar}\left(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}\mathbf{Q}_{\mathbf{R}})\right),\end{aligned}$$

where we use $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}$, $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ *and* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}}$ *to denote the maximum likelihood estimators when the parameters in their subscripts are known. Explicit expressions for the asymptotic variances of* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}$, $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$ *and* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}}$ *are given in the Supplement, Section 3.9.7.*

The first term in the above decomposition, $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}))$, is the same as that in Proposition 3.1, which gave the asymptotic variance if we knew bases for the true envelope. If we set $\mathbf{R} = \mathbf{I}_p$ in the above decomposition, so we are pursuing $\mathbf{Y}$ reduction only, then the decomposition reduces to that given by Cook et al. (2010; Theorem 6.1). Setting $\mathbf{L} = \mathbf{I}_r$, which indicates $\mathbf{X}$ reduction only gives the corresponding result of Cook et al. (2013; Proposition 4.4). The projections $\mathbf{Q}_{\mathbf{R}}$ and $\mathbf{Q}_{\mathbf{L}}$ serve to orthogonalize the random vectors so that the asymptotic variances are additive.

3.4.3 Residual bootstrap

To illustrate the application of the above bootstrap method, we consider a simple model with $p = r = 3$ and $d_X = d_Y = 1$. We generated $\mathbf{R}$, $\mathbf{L}$ and $\boldsymbol{\eta}$ by filling in random numbers from uniform$(0, 1)$ distribution. Then we orthonormalized $\mathbf{R}$, $\mathbf{L}$ and obtained

corresponding $\mathbf{R}_0$, $\mathbf{L}_0$. The covariance matrices were $\mathbf{\Omega} = 5$, $\mathbf{\Omega}_0 = \mathbf{I}_2$, $\mathbf{\Phi} = 1$ and $\mathbf{\Phi}_0 = 10\mathbf{I}_2$. The data vectors $\mathbf{C}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$, $i = 1, \ldots, n$, were simulated from $\mathbf{\Sigma}_{\mathbf{C}}^{1/2}\mathbf{U}_i$ where $\mathbf{U}_i$ is a vector of i.i.d uniform random variables with mean 0 standard deviation 1. Therefore, $\mathbf{C}_i$ would follow a distribution with mean 0, covariance $\mathbf{\Sigma}_{\mathbf{C}}$ and finite fourth moments. A dataset with $n = 100$ observations was generated and $B = 100$ bootstrap datasets were used throughout. We used the bootstrap method to estimate the variances of two estimators: the OLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and the simultaneous envelope estimator $\widehat{\boldsymbol{\beta}}_{1\mathrm{D}}$ which was obtained by using the 1D algorithm without alternating.

Table 3.1 summarizes all the $p \times r = 9$ elements of $\mathrm{vec}(\boldsymbol{\beta})$, $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$ and $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{1\mathrm{D}})$ and their asymptotic, bootstrap and actual standard errors. We included the asymptotic standard errors of the elements in $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which were the square roots of diagonals in $\mathbf{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}/n$ and the asymptotic standard errors of the MLE under normality for the simultaneous envelope model, which is given in Section 3.9, in order to compare with the estimator from the 1D manifold algorithm. We also repeatedly simulated $N = 100$ datasets under the same setting, and estimated the averaged coefficient estimates $\mathrm{E}_N(\widehat{\boldsymbol{\beta}})$ and the actual standard errors of $\mathrm{vec}(\widehat{\boldsymbol{\beta}})$ from the $N = 100$ replicates.

The asymptotic standard errors and the actual standard errors were well estimated by the bootstrap method. Moreover, the estimated asymptotic standard errors for the 1D algorithm are really close to the asymptotic standard errors of the MLE although the normality assumption was violated. As expected, the envelope estimator had much smaller standard errors than that of the OLS estimator.

True parameter									
$\mathrm{vec}(\boldsymbol{\beta}) \times 100$	30	30	32	36	36	38	14	14	15
OLS estimator									
$\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \times 100$	37	37	14	29	42	53	39	2	15
$\mathrm{E}_N\left\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} \times 100$	29	29	32	37	36	38	13	17	13
Asymptotic se$\times 100$	22	22	22	20	20	19	26	26	26
Bootstrap se$\times 100$	21	20	17	17	17	16	25	28	25
Actual se$\times 100$	22	23	22	21	22	18	30	24	23
Simultaneous envelope estimator									
$\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{1\mathrm{D}}) \times 100$	31	34	30	39	43	38	16	18	16
$\mathrm{E}_N\left\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{1\mathrm{D}})\right\} \times 100$	30	30	32	36	35	38	14	14	15
Asymptotic se$\times 100$	3	3	3	3	3	3	2	2	2
Bootstrap se$\times 100$	3	2	3	3	3	3	2	2	2
Actual se$\times 100$	3	2	3	3	3	3	2	2	2

Table 3.1: Bootstrap and estimated asymptotic standard errors of the 9 elements in $\widehat{\boldsymbol{\beta}}$ under the OLS estimator and the simultaneous envelope estimator with fixed $d_X = d_Y = 1$.

3.5 Selection of rank and envelope dimensions

3.5.1 Rank

Since the simultaneous envelope contains the row and the column space of $\boldsymbol{\beta}$, the dimensions d_X and d_Y are bounded below by $d = \text{rank}(\boldsymbol{\beta})$. Thus when determine the dimensions d_X and d_Y, it is helpful to first have some guidance on d. Bura and Cook (2003) developed a chi-squared test for the rank d that requires only that the response variables have finite second moments. The test statistic is $\Lambda_d = n\sum_{j=d+1}^{\min(p,r)} \varphi_j^2$, where $\varphi_1 \geq \cdots \geq \varphi_{\min(p,r)}$ are eigenvalues of the $p \times r$ matrix

$$\widehat{\boldsymbol{\beta}}_{\text{std}} = \{(n-p-1)\}/n\}^{1/2}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\widehat{\boldsymbol{\beta}}_{\text{OLS}}\mathbf{S}_{\mathbf{X}}^{1/2}. \tag{3.5.1}$$

Then they derived that $\boldsymbol{\Lambda}_d$ is asymptotically distributed as a $\chi^2_{(p-d)(r-d)}$ random variable under the null hypothesis that $\text{rank}(\boldsymbol{\beta}) = d$. The rank is then determined by comparing a sequence of test statistics Λ_k, $k = 0, \ldots, \min(p,r) - 1$, to the percentiles of their null distribution $\chi^2_{(p-k)(r-k)}$. The first non-significant value of k will serve as our estimate of the rank of $\boldsymbol{\beta}$.

3.5.2 Envelope dimensions

One way to determine d_X and d_Y is by using a sequence of likelihood ratio tests based on the statistic $\Lambda_{d_X,d_Y} = 2(\widehat{L}_{\text{full}} - \widehat{L}_{d_X,d_Y})$, where $\widehat{L}_{\text{full}} = -(n/2)\log|\mathbf{S_C}| - n(p+r)/2$ is the log likelihood for the standard model and $\widehat{L}_{d_X,d_Y} = -(n/2)\log|\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}}| - (n/2)\text{trace}(\mathbf{S_C}\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}}^{-1})$ is the maximum log likelihood for simultaneous envelope model, where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}} = \mathbf{P}_{\widehat{\mathbf{R}}\oplus\widehat{\mathbf{L}}}\mathbf{S_C}\mathbf{P}_{\widehat{\mathbf{R}}\oplus\widehat{\mathbf{L}}} + \mathbf{Q}_{\widehat{\mathbf{R}}\oplus\widehat{\mathbf{L}}}\mathbf{S_D}\mathbf{Q}_{\widehat{\mathbf{R}}\oplus\widehat{\mathbf{L}}}$ is the MLE of the simultaneous envelope model. Under the null hypothesis, the test statistic Λ_{d_X,d_Y} is asymptotically distributed as a $\chi^2_{(p-d_X)(r-d_Y)}$ random variable. We start testing with $d_X = d_Y = d_0$ where the choice of d_0 can be guided by the Bura-Cook estimator. Then sequentially test $d_X = d_0,\ d_0+1, \ldots$, until first non-significant value, and sequentially test $d_Y = d_0,\ d_0+1, \ldots$, until first non-significant value.

Information criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) could also be used to select the dimension d_X and d_Y, similar to Cook et al. (2010). For dimensions d_X and d_Y, AIC is $\text{A}(d_X, d_Y) = 2k_{d_X,d_Y} - 2\widehat{L}_{d_X,d_Y}$, where $k_{d_X,d_Y} = p(p+3)/2 + r(r+3)/2 + d_X d_Y$ is the number of parameters in the model; and BIC is $\text{B}(d_X, d_Y) = k_{d_X,d_Y}\log(n) - 2\widehat{L}_{d_X,d_Y}$. We search (d_X, d_Y) from (d_0, d_0) to (p, r) and choose the pair that has the smallest AIC or BIC. See Section 3.6.4 for a simulation to illustrate determining the dimensions by likelihood ratio testing, AIC and BIC.

Our experience suggests that BIC is the most favorable criterion for determining dimensions when sample size is not too small. Since the true dimension always exist, BIC is consistent in the sense that the probability of selecting the correct dimension approaches 1. According to Shao (1997) and Yang (2005), BIC would perform better than AIC when

the true model has a simple finite dimensional structure. However, if one is concerned more about the correctness of the envelope model itself over its simplicity, AIC is more favorable since it is more conservative on choosing the dimensions.

In practice, we use cross-validation to determine the optimal dimensions for prediction, although the best predictive dimensions can be different from the true envelope dimensions d_X and d_Y.

3.6 Simulations

In practice, we found the estimator from the 1D algorithm essentially as efficient as the final estimator by alternating because the alternating procedure usually converged after only a few iterations and produced only small changes. To emphasize the 1D algorithm, we use it for all the simulations and examples in this section. In Section 3.6.3, we use a simulation example to demonstrate that estimators from the 1D manifold algorithm may have the same behavior as maximum likelihood estimators.

In Section 3.6.1, we compared the simultaneous envelope estimator to many other methods including OLS, PLS, CCA, RRR, **X**-envelope and **Y**-envelope, where the dimensions for each method were chosen purely based on the simulated data with no prior knowledge. In the Section 3.6.2, the true dimensions were used for each method and we also used two different criterion for prediction and estimation. Given the true dimension, the simultaneous envelope estimators always had the best performance in both prediction and estimation among all estimators and always performed substantially better than OLS.

3.6.1 Prediction with cross-validation

Prediction error of a multivariate linear model is a composite of the variability in $\widehat{\boldsymbol{\beta}}$ and the intrinsic variance in $\boldsymbol{\epsilon}$. As illustrated by Table 3.1, the simultaneous envelope method can substantially increase estimation accuracy in the regression coefficient matrix and thus it may also increase prediction accuracy. We simulated such situations in the following examples where we generated data from joint normal distribution of $\mathbf{X}$ and $\mathbf{Y}$.

We simulated data from two envelope models (E1 and E2) and two reduced-rank regression models (R1 and R2) as follows. The dimensions were $p = 10$, $r = 15$ and varied d, d_X and d_Y. For the envelope models, we generated $\mathbf{R}$, $\mathbf{L}$ and $\boldsymbol{\eta}$ by filling in random numbers from the uniform(0,1) distribution. Then we orthogonalized $\mathbf{R}$ and $\mathbf{L}$ and obtained corresponding $\mathbf{R}_0$ and $\mathbf{L}_0$. For the RRR models, we simulated $\boldsymbol{\beta} = c\mathbf{A}\mathbf{B}^T$ where $c \in \mathbb{R}^1$, $\mathbf{A} \in \mathbb{R}^{r\times d}$, $\mathbf{B} \in \mathbb{R}^{p\times d}$ were filled with uniform$(0, 1)$ random numbers, then $\mathbf{A}$ and $\mathbf{B}$ were standardized to semi-orthogonal matrices. The covariance matrices $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$, $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ were generated as follows.

(E1) Isotropic envelope covariances. $\{\mathbf{\Omega}, \mathbf{\Omega}_0, \mathbf{\Phi}, \mathbf{\Phi}_0\} = \{5\mathbf{I}_{d_X}, \mathbf{I}_{p-d_X}, 5\mathbf{I}_{d_Y}, \mathbf{I}_{r-d_Y}\}$.

(E2) Randomized envelope covariances. All $\mathbf{\Omega}$, $\mathbf{\Omega}_0$, $\mathbf{\Phi}$ and $\mathbf{\Phi}_0$ were randomly generated as $\mathbf{M}\mathbf{M}^T$ where $\mathbf{M}$ had the same size as the corresponding covariances and were filled with uniform$(0, 1)$ numbers.

(R1) Identity covariances. $\mathbf{\Sigma}_{\mathbf{X}} = 5\mathbf{I}_{10}$ and $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}} = 5\mathbf{I}_{15}$.

(R2) Randomized covariances. Both $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ were randomly generated as $\mathbf{M}\mathbf{M}^T$ in the similar way as in model (E2).

We simulated 200 data sets for each setting and used 100 for training and the others for testing. For each method, after obtaining the estimator $\widehat{\boldsymbol{\beta}}$ from one training data set, we used the testing data set to compute the following quantity, which is an estimate of the variance of $\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\mathbf{X}_{\text{new}}$ for a new observation $\mathbf{X}_{\text{new}}$.

$$\boldsymbol{\alpha} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\widehat{\mathbf{\Sigma}}_{\mathbf{X}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \in \mathbb{R}^{r\times r}, \tag{3.6.1}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{X}$ obtained from the testing set. Notice that $\text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}) = \text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}}) + \text{cov}(\boldsymbol{\epsilon})$ for all unbiased estimators, so we could make our comparison in the magnitudes of $\text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}})$. Therefore the Frobenius norm $||\boldsymbol{\alpha}||_F$ was used as a measure of overall prediction error. We also considered the Frobenius norm $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$ as a measure of overall parameter estimation accuracy in the simulation studies in Section 3.6.2.

We used five-fold cross-validation to find the best predictive dimensions for each of the methods. The results were summarized in Table 3.2. Our simulation results suggested that simultaneous envelopes with dimensions determined by cross-validation could be used in practice as a powerful prediction method. The simultaneous envelope methods had significantly better performance than all the other methods in most cases. For each of the other methods, it could perform as well as the simultaneous envelope in some cases but could have much worse performances in other cases.

In model (R2), where the true envelope dimensions were $(d_X, d_Y) = (p, r)$, the envelope estimators with its true dimensions will be equivalent to OLS. Nevertheless, the $\mathbf{X}$-envelope and the simultaneous envelope showed significant improvement over OLS in prediction by using the best predictive dimensions. This was probably due to the variance-bias tradeoff in finite samples. Model (R1) is both an envelope model and a RRR model with $d = d_X = d_Y$. It has the simplest structure, but the simultaneous envelope estimators still produced significant improvement over OLS. RRR outperformed OLS but was still not as efficient as the simultaneous envelope estimators. Model (E2) produced immaterial covariances $\mathbf{\Phi}_0$ and $\mathbf{\Omega}_0$ that had much larger eigenvalues than the material parts $\mathbf{\Phi}$ and $\mathbf{\Omega}$. The simultaneous envelope method efficiently discarded the immaterial information and

performed drastically better than other methods in models (E2). Moreover, CCA aimed for the most correlated pairs of $\widehat{\mathbf{R}}^T\mathbf{X}$ and $\widehat{\mathbf{L}}^T\mathbf{Y}$ and mistakenly found the immaterial pairs in model (E2).

(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	$\mathbf{X}$-env.	$\mathbf{Y}$-env.	S. env.	S. E.$\leq$
E1	$(3,3,100)$	3.15	1.55	**1.14**	2.46	1.50	2.53	**0.63**	0.06
	$(3,3,400)$	0.70	0.46	**0.22**	2.23	0.46	0.58	**0.12**	0.01
	$(3,3,900)$	0.30	0.14	**0.08**	2.28	0.14	0.25	**0.05**	0.006
E2	$(2,5,400)$	0.82	0.40	2.17	0.28	**0.22**	0.47	**0.17**	0.04
	$(5,3,400)$	1.00	0.51	1.76	**0.36**	0.55	0.65	**0.27**	0.07
	$(7,4,400)$	1.01	0.82	1.75	0.68	0.75	**0.59**	**0.32**	0.08
R1	$(1,200)$	1.68	**0.85**	1.01	1.05	0.97	1.53	**0.73**	0.03
	$(2,200)$	1.68	0.98	1.49	1.89	**0.97**	1.52	**0.97**	0.04
	$(3,200)$	1.68	1.35	2.00	2.38	**1.19**	1.52	**1.18**	0.05
R2	$(2,200)$	2.63	1.23	8.94	**0.90**	1.08	2.63	**1.08**	0.11*
	$(3,200)$	2.86	1.54	26.6	1.52	**1.08**	2.79	**1.08**	0.15*
	$(4,200)$	2.96	2.67	16.4	1.97	**1.45**	2.96	**1.45**	0.13*

Table 3.2: Prediction performances measured by $||\boldsymbol{\alpha}||_F$, where * in the last column means the corresponding S.E. upper bound were computed without the S.E. of CCA. The best two methods for each setting were emphasized with boldface.

To further study how the relative magnitude of immaterial variation over material variation can affect the performance of envelope methods, we repeated the above simulations in model (E2) for varying $d = d_X = d_Y$ from 1 to 8 with the same sample size $N = 200$. The true envelope dimensions were used instead of using cross-validation. We then plotted the averaged prediction error $||\boldsymbol{\alpha}||_F$ over that of OLS in Figure 3.2. When the rank of $\boldsymbol{\beta}$ was smaller than 5, which is the half of its maximum possible rank, the immaterial variance $\boldsymbol{\Phi}_0$ contained the larger eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ hence the $\mathbf{Y}$-envelope and the simultaneous envelope estimators had drastic gains in efficiency over OLS estimators. With increased rank of $\boldsymbol{\beta}$, the immaterial variance parts were reduced but the simultaneous envelope reduction can still gain significantly over the OLS estimator. Even for $d = 8$, the prediction error measurement $||\boldsymbol{\alpha}||_F$ of the simultaneous envelope estimator was only 70% of that of the OLS estimator.

3.6.2 Knowing the true dimensions

We used the same models as in Section 3.6.1. We simulated 1000 data sets for each setting and used 500 for training and the others for testing. We used the known dimension for each method: the true ranks of $\boldsymbol{\beta}$ were used for CCA and RRR, the $\mathbf{X}$-envelope dimensions d_X were used for PLS, and the true envelope dimensions were used for envelope methods. The Frobenius norm $||\boldsymbol{\alpha}||_F$ was used as a measure of overall prediction error. And we also used the Frobenius norm $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$ as a measure of overall parameter estimation accuracy.

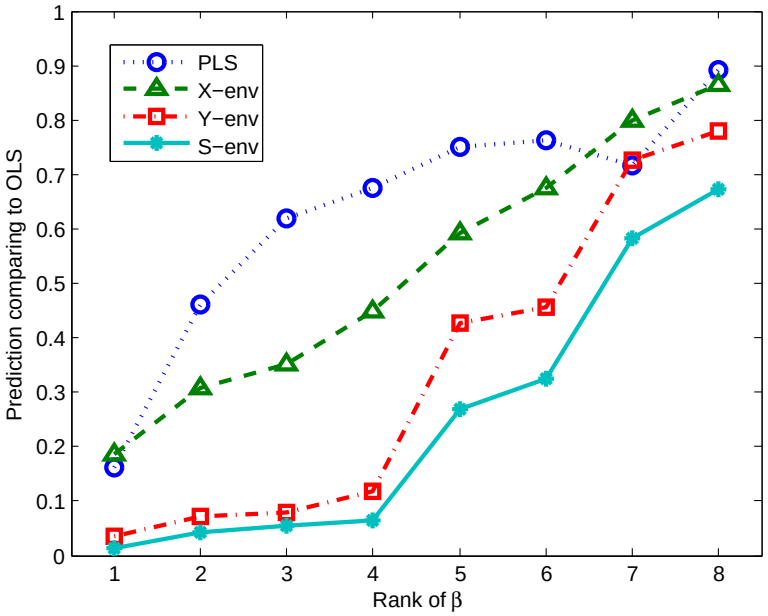


Figure 3.2: Prediction versus OLS for varying $d = d_X = d_Y$. Each point is the ratio of $||\boldsymbol{\alpha}||_F$ for a specified estimator and for OLS estimator. The lines for CCA and RRR (not shown in the plot) were always larger than or close to 1.

Simulation results were summarized in Table 3.3. We tried different dimensions and different sample sizes for all these models. As expected, the simultaneous envelope estimators always had the best performance among all estimators and always performed substantially better than OLS. For each of the other methods, it could perform as well as the simultaneous envelope in some cases but could have much worse performances in other cases. We did not include the model (R2) because the true envelope dimensions were p and r, which gave no reductions. Interpretations of the results were similar as in Section 3.6.

3.6.3 Performance of the 1D algorithm

The simulation results in Section 3.6.1 were all based on the 1D algorithm. The full Grassmann manifold optimization, which gives the MLE solution, would give indistinguishable performances for small dimensions but could have trouble converging in larger dimensions. To gain some insight on the computation cost of 1D algorithm versus the full Grassmann manifold optimization, we simulated 100 data sets from model (E2) with $N = 400$ and computed the averaged CPU time for estimating an envelope using the two methods. When $d_X = d_Y$ increased from 2 to 3, the ratio of the averaged time costs on full Grassmannian optimization over that of the 1D algorithm were increasing from 1.1 to 2.8 for estimating $\mathbf{X}$-envelopes and from 1.4 to 3.8 for estimating $\mathbf{Y}$-envelopes.

To examine how well the 1D manifold algorithm estimators can approximate the MLE,

Prediction Criterion $\|\|\boldsymbol{\alpha}\|\|_F$									
(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	**X**-env.	**Y**-env.	S. env.	S. E.$\leq$
E1	$(3, 3, 100)$	3.2	1.3	0.75	2.3	1.3	**0.57**	**0.48**	0.02
	$(3, 3, 400)$	0.70	0.32	0.14	0.37	0.31	**0.11**	**0.10**	0.004
	$(3, 3, 900)$	0.30	0.14	0.06	0.15	0.13	**0.05**	**0.04**	0.002
E2	$(2, 5, 200)$	1.6	0.55	2.8	1.6	**0.51**	0.76	**0.28**	0.03
	$(5, 3, 200)$	1.9	1.2	11	1.9	1.2	**0.12**	**0.10**	0.04*
	$(7, 4, 200)$	1.9	1.8	44	1.9	1.5	**0.29**	**0.23**	0.04*
R1	$(1, 100)$	3.8	**1.5**	1.83	2.3	2.2	1.7	**1.4**	0.03
	$(3, 100)$	3.8	**2.5**	3.9	2.7	2.8	2.7	**2.6**	0.03
	$(5, 100)$	3.8	**3.0**	4.7	3.3	3.3	3.2	**3.1**	0.04
Estimation Criterion $\|\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|\|_F$									
(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	**X**-env.	**Y**-env.	S. env.	S. E.$\leq$
E1	$(3, 3, 100)$	253	28	31	142	28	**25**	**11**	2.6
	$(3, 3, 400)$	14	2	**1**	3	2	**1**	**1**	0.1
	$(3, 3, 900)$	2.6	0.3	**0.2**	0.4	0.3	**0.2**	**0.1**	0.02
E2	$(2, 5, 200)$	451	**23**	116	425	127	197	**59**	11
	$(5, 3, 200)$	994	94	80	946	493	**45**	**26**	26
	$(7, 4, 200)$	1152	599	225	1112	876	**154**	**104**	36
R1	$(1, 100)$	73	**30**	36	46	42	34	**27**	0.7
	$(3, 100)$	73	**49**	76	53	54	52	**50**	0.8
	$(5, 100)$	73	**58**	90	64	64	62	**60**	0.8

Table 3.3: Prediction performances measured by $||\boldsymbol{\alpha}||_F$ and estimation performances measured by $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$, where * in the last column means the corresponding S.E. upper bound were computed without the S.E. of CCA. The best two methods for each setting were emphasized with boldface.

we studied probability plots of the asymptotic χ^2 likelihood ratio test statistics. Using model (E2), we simulated 100 dataset with sample size $N = 200$ with $p = r = 7$ and $d_X = d_Y = 3$. Consider the following hypothesis test given d_X and d_Y,

$$H_0 : \mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) \oplus \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta}) = \text{span}(\mathbf{R} \oplus \mathbf{L});$$
$$H_a : \mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T) \oplus \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\boldsymbol{\beta}) \neq \text{span}(\mathbf{R} \oplus \mathbf{L}).$$

The likelihood ratio test statistic for this hypothesis is $\mathbf{\Lambda} = -2\log(L(\mathbf{R}, \mathbf{L})) + 2\log(L(\widehat{\mathbf{R}}, \widehat{\mathbf{L}}))$, where $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{L}}$ are MLEs, is asymptotically χ^2 distributed. The degrees of freedom is the sum of the two Grassmann manifolds dimensions: $(p - d_X)d_X + (r - d_Y)d_Y = 24$. We computed the likelihood ratio test statistics using the 1D algorithm to estimate $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{L}}$. Figure 3.3 summarized the calculated sample quantiles of the one hundred statistics and the expected quantiles of the χ^2_{24} distribution. All the points fall near a straight line, indicating that the estimators from the 1D algorithm behave as MLEs.

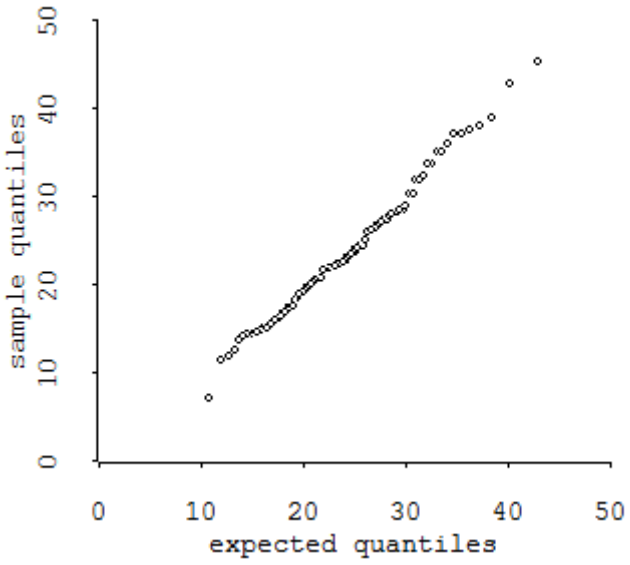


Figure 3.3: Probability plot based on the 1D algorithm

3.6.4 Determining the envelope dimensions

We simulated 100 datasets for each of the following settings, and performed the chi-square test for the rank of $\boldsymbol{\beta}$ with level 0.05, the likelihood ratio test with level 0.01, AIC model selection and BIC model selection. We used model (E2) in Section 3.6.1 with $p = 6$ predictors and $r = 8$ responses in two settings: (I) $N = 100$ observations for each dataset with $\mathbf{X}$- and $\mathbf{Y}$-envelope dimensions $d_X = d_Y = 2$. (II) $N = 300$ observations for each dataset with $\mathbf{X}$- and $\mathbf{Y}$-envelope dimensions $d_X = 2$, $d_Y = 3$. Table 3.4 summarizes

the simulation performances of various tests. Clearly, the χ^2 test quite often reveals the correct dimension of $\boldsymbol{\beta}$ and BIC has the best performance on determining the envelope dimensions.

We further studied the likelihood ratio test statistics obtained by the 1D algorithm and SIMPLS. We used $p = r = 7$, $d_X = d_Y = 3$ and simulated 100 dataset with $N = 300$ observations from model (E2) in Section 3.6.1. The likelihood ratio test was computed for the hypothesis $d_X = d_Y = 3$. The test statistic follows a χ^2-distribution with $pr - d_X d_Y = 40$ degrees of freedom as shown in Section 3.5.2. The probability plot of expected quantiles versus sample quantiles from the 100 datasets are given in Figure 3.4. Since the test statistic depends on our estimators of $\mathbf{R}$ and $\mathbf{L}$, the probability plot should follow a straight line if the estimators behave as MLEs. From this plot, we see again that the behavior of the 1D algorithm is as expected for MLEs. However, the SIMPLS algorithm produced noticeable curvature in its probability plot.

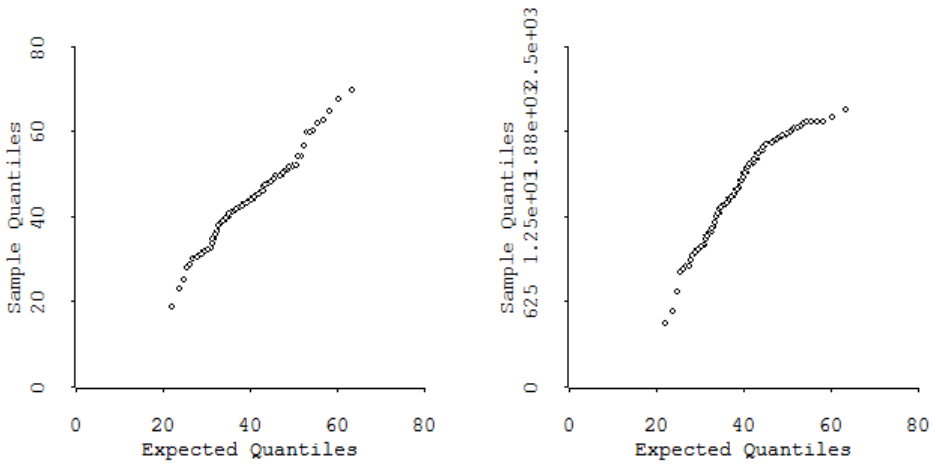


Figure 3.4: Chi-squared distribution probability plots. The 1D algorithm results the left plot, and the SIMPLS algorithm results the right plot. The reference distribution is the χ^2_{40}-distribution.

3.7 Biscuit NIR spectroscopy data

This experiment involved varying the composition of biscuit dough. Two sets of dough pieces were measured, a calibration set and a prediction set. They were created and measured as two distinct sets, on separate occasions, and do not result from a random (or any other) split of a larger set. We have $r = 4$ response variables: fat, sucrose, flour and water percentages of biscuit dough. To keep the number of predictors less than the sample size of the training set, we used the spectral range 1200-1280nm with 4nm steps, which gave $p = 20$ predictors with $n_1 = 39$ training samples and $n_2 = 31$ testing samples.

	correctness on	setting (I)	setting (II)
χ^2	d	89%	97%
LRT	d_X	98%	97%
	d_Y	100%	90%
	d_X and d_Y	98%	89%
AIC	d_X	75%	92%
	d_Y	67%	63%
	d_X and d_Y	56%	61%
BIC	d_X	98%	99%
	d_Y	100%	100%
	d_X and d_Y	98%	99%

Table 3.4: The numbers indicate how many times each true/estimated dimensions occurs among the 100 datasets.

The prediction accuracy was based on the sum of squared error for predicting the given testing set.

The rank test of $\boldsymbol{\beta}$ indicated that $d = \text{rank}(\boldsymbol{\beta}) = 2$ at the 0.05 significant level. Simultaneous envelope dimension $d_Y = 2$ was selected by AIC, BIC and LRT; $d_X = 10, 3, 8$ was selected by AIC, BIC and LRT respectively. Based on the prediction error, AIC chose the best dimension, while BIC made a bad choice, which was probably due to the small sample size of the training set.

The sum of squared prediction errors (prediction SSE) on the testing sets are summarized in the Figure 3.5. When the number of components was less than 8, the envelope and SIMPLS estimators had similar performances and were both inferior to OLS because the dimension was too small to cover all the material information. Starting at 8 components, the envelope estimator performed much better than the SIMPLS and OLS estimators. It converged to the OLS estimator at 13 components and stayed the same thereafter. However, SIMPLS performed much worse than OLS unless the number of component was larger than 16. To aid visualization, prediction errors from CCA and RRR were not plotted because they were more than twice of that of OLS. We also considered predictions based on $\mathbf{X}$- or $\mathbf{Y}$-envelope reduction alone. The prediction errors for the $\mathbf{X}$-envelope estimator traced the errors from the simultaneous envelope estimator but were uniformly larger. The prediction errors for the $\mathbf{Y}$-envelope estimators were indistinguishable to that of the OLS estimator for $d_Y > 1$ and were larger than that for $d_Y = 1$.

3.8 Discussion

The partial envelope model (Su and Cook, 2011) is an extension of the envelope model in which some predictors are of special interest. Suppose $\mathbf{X}$ can be partitioned into $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$, and the regression coefficient becomes $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)^T$

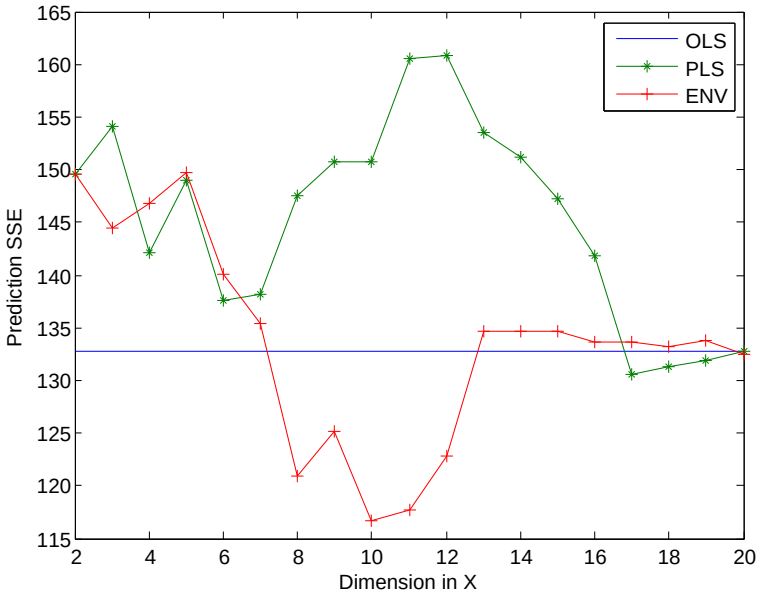


Figure 3.5: Prediction sum of square errors on the testing set. X-axis denote the numbers of components for PLS and simultaneous envelope **X**-dimension, d_X, where the **Y**-dimension of simultaneous envelope is fixed at $d_Y = 2$. To help visualization, the CCA performance is not included and the SSE for CCA are all greater than OLS no matter how many components to use.

correspondingly. The partial envelope model focus on estimating $\boldsymbol{\beta}_1$ instead of $\boldsymbol{\beta}$, and can be written as,

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}_1(\mathbf{X}_1 - \boldsymbol{\mu}_{\mathbf{X}1}) + \boldsymbol{\beta}_2(\mathbf{X}_2 - \boldsymbol{\mu}_{\mathbf{X}2}) + \boldsymbol{\epsilon}, \tag{3.8.1}$$

where $\boldsymbol{\mu}_{\mathbf{X}1} = \mathrm{E}(\mathbf{X}_1)$ and $\boldsymbol{\mu}_{\mathbf{X}2} = \mathrm{E}(\mathbf{X}_2)$. The partial envelope in this model is $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, which is contained in the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$. Since the partial envelope model focuses on a smaller subspace, there is potentially more efficiency gain than with the envelope model.

It is straightforward to extent partial envelope to simultaneous envelopes. Let $\boldsymbol{\Sigma}_{\mathbf{X}1} \in \mathbb{R}^{p_1 \times p_1}$ be the covariance matrix of $\mathbf{X}_1$, and let $\boldsymbol{\Sigma}_{\mathbf{X}2} \in \mathbb{R}^{p_2 \times p_2}$ be the covariance matrix of $\mathbf{X}_2$. We can then study the simultaneous partial envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}_1}}(\boldsymbol{\beta}_1^T)$ to reduce $\mathbf{Y}$ and to reduce $\mathbf{X}_1$.

The simultaneous envelope model contains the envelope models in Cook et al. (2010) and Cook et al. (2013) as special cases. Extensions to the simultaneous envelope methods provide unified treatments for these methods. For instance, variable selection in multivariate linear regression has started to become an interesting and important research area. Multivariate linear methods that have well performances in classical problems were extended to high-dimensional situations. To deal with high-dimensional problems where $p, r > n$, Chun and Keles (2010) introduced sparse PLS and Chen and Huang (2012) proposed sparse RRR. We conjecture that an variable selection extension of the simultaneous envelope methodology will have nice applications in high-dimensional settings. Alternatively, we can first apply the SIMPLS algorithm, which is actually an envelope algorithm shown by Cook et al. (2013), to reduce the dimensionality and then apply the simultaneous envelope methods to gain efficiency in estimation.

3.9 Proofs

3.9.1 Lemma 3.1 and Lemma 3.2

Lemma 3.1 and Lemma 3.2 follow directly from the discussions in Section 3.2 and the definition of envelopes.

3.9.2 Proposition 3.1

Proof. Let $\mathbb{X} \in \mathbb{R}^{n \times p}$ and $\mathbb{Y} \in \mathbb{R}^{n \times r}$ be the centered data matrices from i.i.d. random vectors $\mathbf{X}$ and $\mathbf{Y}$. Then $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}} = \mathbf{L}\mathbf{L}^T\mathbb{Y}^T\mathbb{X}\mathbf{R}(\mathbf{R}^T\mathbb{X}^T\mathbb{X}\mathbf{R})^{-1}\mathbf{R}^T$.

For the variance of $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$, we decompose it into two terms, conditioning on $\mathbb{X}$.

$$\begin{aligned}
\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) &= \mathrm{var}\left\{\mathrm{vec}(\mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1})\right\} \\
&= \mathrm{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)\right\} \\
&= \mathrm{var}\left\{\mathrm{E}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X}\right\}\right\} \\
&\quad +\mathrm{E}\left\{\mathrm{var}(\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X})\right\} \\
&\equiv VE + EV.
\end{aligned}$$

The first term is

$$\begin{aligned}
VE &= \mathrm{var}\left\{\mathrm{E}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X}\right\}\right\} \\
&= \mathrm{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{E}\{\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X}\}\right\} \\
&= \mathrm{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\boldsymbol{\beta}\mathbb{X}^T)\right\} \\
&= \mathrm{var}\left\{\mathrm{vec}(\boldsymbol{\beta}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1})\right\} = 0.
\end{aligned}$$

Then the variance is just the second term which is

$$\begin{aligned}
\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) &= EV = \mathrm{E}\left\{\mathrm{var}(\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X})\right\} \\
&= \mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\mathrm{var}(\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X})\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)^T\} \\
&= \mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)\left(\mathbf{I}_n\otimes\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\right)\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\otimes\mathbf{I}_r\right)^T\} \\
&= \mathrm{E}\{(\mathbb{X}^T\mathbb{X})^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\} \\
&= f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}},
\end{aligned}$$

where the last equality come from the fact that $\mathbb{X}^T\mathbb{X}\sim W_p(\boldsymbol{\Sigma}_{\mathbf{X}},n-1)$. The degrees of freedom $n-1$ are greater than $p-1$. Then $(\mathbb{X}^T\mathbb{X})^{-1}$ follows an inverse Wishart distribution $W_p^{-1}(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1},n-1)$, and its mean equals $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}/(n-p-2)=f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$.

Similarly for $\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}$, we have the following expressions by noticing that $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}}) = (\mathbf{R}\oplus\mathbf{L})\mathrm{vec}(\widehat{\boldsymbol{\eta}})$ and $\widehat{\boldsymbol{\eta}}$ is the OLS estimator of $\mathbf{L}^T\mathbf{Y}$ on $\mathbf{R}^T\mathbf{X}$.

$$\begin{aligned}
\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})) &= f_x^{-1}(\mathbf{R}\otimes\mathbf{L})(\boldsymbol{\Sigma}_{\mathbf{R}^T\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{L}^T\mathbf{Y}|\mathbf{R}^T\mathbf{X}})(\mathbf{R}^T\otimes\mathbf{L}^T) \\
&= f_x^{-1}(\mathbf{R}\otimes\mathbf{L})((\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R})^{-1}\otimes(\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}))(\mathbf{R}^T\otimes\mathbf{L}^T) \\
&= f_x^{-1}\left(\mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R})^{-1}\mathbf{R}^T\right)\otimes\left(\mathbf{L}\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}\mathbf{L}^T\right) \\
&= f_x^{-1}\left(\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T\right)\otimes\left(\mathbf{P}_{\mathbf{L}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{L}}\right).
\end{aligned}$$

Next, by noticing that

$$\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} = \mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R})^{-1}\mathbf{R}^T+\mathbf{R}_0(\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}_0)^{-1}\mathbf{R}_0^T, \tag{3.9.1}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{P}_{\mathbf{L}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{L}}+\mathbf{P}_{\mathbf{L}_0}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{L}_0}, \tag{3.9.2}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T+\mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T, \tag{3.9.3}$$

we have

$$\begin{aligned}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) &= f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \\ &= f_x f_p^{-1}\text{var}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})) + f_p^{-1}\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T \\ &\quad + f_p^{-1}\mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T + f_p^{-1}\mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T,\end{aligned}$$

where $\boldsymbol{\Omega} = \mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}$, $\boldsymbol{\Omega}_0 = \mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}_0$, $\boldsymbol{\Phi}_0 = \mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}_0$ and $\boldsymbol{\Phi} = \mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{L}$. Then the result follows directly. □

3.9.3 Lemma 3.3

From Proposition 2.2 in Cook et al. (2010), we have $\mathcal{E}_{\mathbf{M}_i}(\mathcal{S}_i) = \sum_{j=1}^{q_i}\mathbf{P}_{ij}\mathcal{S}_i$, $i = 1, 2$, where $\mathbf{P}_{ij}$ is the projection onto the j-th eigenspace of $\mathbf{M}_i$. Since $\mathbf{M} = \mathbf{M}_1 \oplus \mathbf{M}_2$, then the eigen-projections of $\mathbf{M}$ will be $\mathbf{P}_{1j} \oplus \mathbf{0}_{p_2}$, $j = 1, \ldots, q_1$, and $\mathbf{0}_{p_1} \oplus \mathbf{P}_{2k}$, $k = 1, \ldots, q_2$. Therefore, applying Proposition 2.2 of Cook et al. (2010) again, we have

$$\begin{aligned}\mathcal{E}_{\mathbf{M}}(\mathcal{S}) &= \left\{\sum_{j=1}^{q_1}[(\mathbf{P}_{1j} \oplus \mathbf{0})(\mathcal{S}_1 \oplus \mathcal{S}_2)]\right\} + \left\{\sum_{k=1}^{q_2}[(\mathbf{0} \oplus \mathbf{P}_{2k})(\mathcal{S}_1 \oplus \mathcal{S}_2)]\right\} \\ &= \{\mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1) \oplus \mathbf{0}\} + \{\mathbf{0} \oplus \mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2)\} \\ &= \mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1) \oplus \mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2).\end{aligned}$$

3.9.4 Lemma 3.4

The log-det term

In this section, we compute the log-determinant term of $\log|\boldsymbol{\Sigma}_{\mathbf{C}}|$. Apply the following orthogonal transformation to $\boldsymbol{\Sigma}_{\mathbf{C}}$,

$$\begin{aligned}\mathbf{O}^T\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{O} &= \begin{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{R}_0 \end{pmatrix}^T & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \mathbf{L} & \mathbf{L}_0 \end{pmatrix}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{XY}}^T & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix} \begin{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{R}_0 \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} \mathbf{L} & \mathbf{L}_0 \end{pmatrix} \end{pmatrix} \\ &= \begin{pmatrix} \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{pmatrix} & \begin{pmatrix} \boldsymbol{\Omega}\boldsymbol{\eta}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ \begin{pmatrix} \boldsymbol{\eta}\boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} & \begin{pmatrix} \boldsymbol{\Phi} + \boldsymbol{\eta}\boldsymbol{\Omega}\boldsymbol{\eta}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_0 \end{pmatrix} \end{pmatrix}.\end{aligned}$$

We can further transform it into a block diagonal matrix as in the following general case:

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{pmatrix}. \quad (3.9.4)$$

By taking the transformation matrix $\mathbf{L} = \begin{pmatrix} \mathbf{I} & \mathbf{L}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ where $\mathbf{L}_{12} = \begin{pmatrix} -\boldsymbol{\eta}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, we have the block diagonal matrix

$$\mathbf{L}^T\mathbf{O}^T\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{O}\mathbf{L} = \text{diag}\{\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Phi}, \boldsymbol{\Phi}_0\} \equiv \boldsymbol{\Sigma}_{\mathbf{diag}}. \tag{3.9.5}$$

Also, because of the determinants of $\mathbf{L}$ and $\mathbf{O}$ are both 1, $|\boldsymbol{\Sigma}_{\mathbf{C}}| = |\boldsymbol{\Sigma}_{\mathbf{diag}}|$, and

$$\log|\boldsymbol{\Sigma}_{\mathbf{C}}| = \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \log|\boldsymbol{\Phi}| + \log|\boldsymbol{\Phi}_0|. \tag{3.9.6}$$

The trace term

In this section, we compute the trace term of $\text{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1})$.

$$\text{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}) = \text{trace}(\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}\mathbf{L}^T\mathbf{O}^T) = \text{trace}(\mathbf{L}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}).$$

Write $\mathbf{O} \equiv \begin{pmatrix} \mathbf{O}_g & 0 \\ 0 & \mathbf{O}_h \end{pmatrix}$, where $\mathbf{O}_g = (\mathbf{R}, \mathbf{R}_0)$ and $\mathbf{O}_h = (\mathbf{L}, \mathbf{L}_0)$. Next, we compute the matrix

$$\mathbf{L}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L} = \mathbf{L}^T \begin{pmatrix} \mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g & \mathbf{O}_g^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h \\ \mathbf{O}_h^T\mathbf{S}_{\mathbf{XY}}^T\mathbf{O}_g & \mathbf{O}_h^T\mathbf{S}_{\mathbf{Y}}\mathbf{O}_h \end{pmatrix} \mathbf{L} = \begin{pmatrix} \mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g & * \\ * & \mathbf{M}_1 \end{pmatrix}.$$

Since we only need the trace of $\mathbf{L}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}$, and $\boldsymbol{\Sigma}_{\mathbf{diag}}$ is block diagonal, so the off-diagonal blocks of $\mathbf{L}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L}$ are not needed. And the matrix $\mathbf{M}_1$ can be computed to be

$$\mathbf{M}_1 = \mathbf{O}_h^T\mathbf{S}_{\mathbf{Y}}\mathbf{O}_h + (\mathbf{O}_g\mathbf{L}_{12})^T\mathbf{S}_{\mathbf{X}}(\mathbf{O}_g\mathbf{L}_{12}) + (\mathbf{O}_g\mathbf{L}_{12})^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h + \left[(\mathbf{O}_g\mathbf{L}_{12})^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h\right]^T.$$

Therefore,

$$\begin{aligned} \text{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}) &= \text{trace}(\mathbf{L}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{L}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}) \\ &= \text{trace}(\mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g\text{diag}(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Omega}_0^{-1})) + \text{trace}(\mathbf{M}_1\text{diag}(\boldsymbol{\Phi}^{-1}, \boldsymbol{\Phi}_0^{-1})) \\ &\equiv \text{trace}_1 + \text{trace}_2. \end{aligned}$$

The first trace term,

$$\begin{aligned} \text{trace}_1 &= \text{trace}(\mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g\text{diag}(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Omega}_0^{-1})) \\ &= \text{trace}\left\{ \begin{pmatrix} \mathbf{R}^T \\ \mathbf{R}_0^T \end{pmatrix} \mathbf{S}_{\mathbf{X}} \begin{pmatrix} \mathbf{R} & \mathbf{R}_0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}^{-1} & 0 \\ 0 & \boldsymbol{\Omega}_0^{-1} \end{pmatrix} \right\} \\ &= \text{trace}\{\mathbf{R}^T\mathbf{S}_{\mathbf{X}}\mathbf{R}\boldsymbol{\Omega}^{-1}\} + \text{trace}\{\mathbf{R}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\} = \text{trace}\{\mathbf{S}_{\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\}. \end{aligned}$$

Similarly for the second trace term,

$$\begin{aligned} \text{trace}_2 &= \text{trace}(\mathbf{M}_1\text{diag}(\boldsymbol{\Phi}^{-1}, \boldsymbol{\Phi}_0^{-1})) \\ &= \text{trace}\{\mathbf{L}^T\mathbf{S}_{\mathbf{Y}}\mathbf{L}\boldsymbol{\Phi}^{-1}\} + \text{trace}\{\mathbf{L}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{L}_0\boldsymbol{\Phi}_0^{-1}\} \\ &\quad + \text{trace}\{\boldsymbol{\eta}\mathbf{R}^T\mathbf{S}_{\mathbf{X}}\mathbf{R}\boldsymbol{\eta}^T\boldsymbol{\Phi}^{-1}\} - 2 \times \text{trace}\{\boldsymbol{\eta}\mathbf{R}^T\mathbf{S}_{\mathbf{XY}}\mathbf{L}\boldsymbol{\Phi}^{-1}\}. \end{aligned}$$

Partial derivatives of the objective function

The objective function $\mathrm{F}(\mathbf{\Sigma_C})$ now becomes an objective function of all the parameters

$$\begin{aligned}\mathrm{F}(\mathbf{R},\mathbf{L},\boldsymbol{\eta},\mathbf{\Phi},\mathbf{\Phi}_0,\mathbf{\Omega},\mathbf{\Omega}_0) &= \log|\mathbf{\Omega}|+\log|\mathbf{\Omega}_0|+\log|\mathbf{\Phi}|+\log|\mathbf{\Phi}_0| \\ &\quad +\mathrm{trace}\{\mathbf{R}^T\mathbf{S_X}\mathbf{R}\mathbf{\Omega}^{-1}\}+\mathrm{trace}\{\mathbf{R}_0^T\mathbf{S_X}\mathbf{R}_0\mathbf{\Omega}_0^{-1}\} \\ &\quad +\mathrm{trace}\{\mathbf{L}^T\mathbf{S_Y}\mathbf{L}\mathbf{\Phi}^{-1}\}+\mathrm{trace}\{\mathbf{L}_0^T\mathbf{S_Y}\mathbf{L}_0\mathbf{\Phi}_0^{-1}\} \\ &\quad +\mathrm{trace}\{\boldsymbol{\eta}\mathbf{R}^T\mathbf{S_X}\mathbf{R}\boldsymbol{\eta}^T\mathbf{\Phi}^{-1}\}-2\mathrm{trace}\{\mathbf{R}\boldsymbol{\eta}\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L}\mathbf{\Phi}^{-1}\}.\end{aligned}$$

We will apply the following result repeatedly in this section.

$$\arg\min_{\mathbf{A}}\{\log|\mathbf{A}|+\mathrm{trace}(\mathbf{A}^{-1}\mathbf{B})\}=\mathbf{B}, \tag{3.9.7}$$

where $\mathbf{B}$ is positive definite symmetric matrix and the minimization is over all positive definite symmetric matrices.

Partial minimization In this section, we minimize the objective function over $\mathbf{\Omega}$, $\mathbf{\Omega}_0$, $\mathbf{\Phi}$ and $\mathbf{\Phi}_0$. Applying (3.9.7), we have

$$\widehat{\mathbf{\Omega}}=\widehat{\mathbf{R}}^T\mathbf{S_X}\widehat{\mathbf{R}}; \tag{3.9.8}$$

$$\widehat{\mathbf{\Omega}}_0=\widehat{\mathbf{R}}_0^T\mathbf{S_X}\widehat{\mathbf{R}}_0; \tag{3.9.9}$$

$$\widehat{\mathbf{\Phi}}_0=\widehat{\mathbf{L}}_0^T\mathbf{S_Y}\widehat{\mathbf{L}}_0. \tag{3.9.10}$$

After we obtain the minimizer of $\widehat{\boldsymbol{\eta}}$ and substitute it into the objective function, we can also use (3.9.7) to get the following.

$$\widehat{\mathbf{\Phi}}=\widehat{\mathbf{L}}^T\left(\mathbf{S_Y}-\mathbf{S_{XY}^T}\widehat{\mathbf{R}}\left(\widehat{\mathbf{R}}^T\mathbf{S_X}\widehat{\mathbf{R}}\right)^{-1}\widehat{\mathbf{R}}^T\mathbf{S_{XY}}\right)\widehat{\mathbf{L}}. \tag{3.9.11}$$

We next compute the partial derivative with respect to $\boldsymbol{\eta}$.

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\eta}}\mathrm{F}(\mathbf{R},\mathbf{L},\boldsymbol{\eta},\mathbf{\Phi},\mathbf{\Phi}_0,\mathbf{\Omega},\mathbf{\Omega}_0) &= \frac{\partial}{\partial\boldsymbol{\eta}}\left(\mathrm{trace}\{\boldsymbol{\eta}\mathbf{R}^T\mathbf{S_X}\mathbf{R}\boldsymbol{\eta}^T\mathbf{\Phi}^{-1}\}\right) \\ &\quad -2\frac{\partial}{\partial\boldsymbol{\eta}}\left(\mathrm{trace}\{\boldsymbol{\eta}\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L}\mathbf{\Phi}^{-1}\}\right) \\ &= 2\mathbf{R}^T\mathbf{S_X}\mathbf{R}\boldsymbol{\eta}^T\mathbf{\Phi}^{-1}-2\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L}\mathbf{\Phi}^{-1}.\end{aligned}$$

Set it to 0, we have

$$\widehat{\boldsymbol{\eta}}^T=(\widehat{\mathbf{R}}^T\mathbf{S_X}\widehat{\mathbf{R}})^{-1}(\widehat{\mathbf{R}}^T\mathbf{S_{XY}}\widehat{\mathbf{L}}). \tag{3.9.12}$$

Substituting (3.9.8), (3.9.9), (3.9.10) and (3.9.12) into the full objective function, we have the following partially optimized objective function

$$\begin{aligned}\mathrm{F}(\mathbf{R},\mathbf{L},\mathbf{\Phi}) &= \log|\mathbf{R}^T\mathbf{S_X}\mathbf{R}| + \log|\mathbf{R}_0^T\mathbf{S_X}\mathbf{R}_0| + \log|\mathbf{\Phi}| + \log|\mathbf{L}_0^T\mathbf{S_Y}\mathbf{L}_0| \\ &\quad +\mathrm{trace}\{\mathbf{I}_l\} + \mathrm{trace}\{\mathbf{I}_{p-l}\} + \mathrm{trace}\{\mathbf{L}^T\mathbf{S_Y}\mathbf{L}\mathbf{\Phi}^{-1}\} + \mathrm{trace}\{\mathbf{I}_{r-m}\} \\ &\quad +\mathrm{trace}\{(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})^T(\mathbf{R}^T\mathbf{S_X}\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})\mathbf{\Phi}^{-1}\} \\ &\quad -2\times\mathrm{trace}\{(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})^T(\mathbf{R}^T\mathbf{S_X}\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})\mathbf{\Phi}^{-1}\}.\end{aligned}$$

If we ignore the terms that do not change with $\mathbf{\Phi}$, the rest part of the objective function becomes

$$\mathrm{F}(\mathbf{\Phi}) = \log|\mathbf{\Phi}| + \mathrm{trace}\{\mathbf{L}^T\mathbf{S_Y}\mathbf{L}\mathbf{\Phi}^{-1}\} - \mathrm{trace}\{(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})^T(\mathbf{R}^T\mathbf{S_X}\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{S_{XY}}\mathbf{L})\mathbf{\Phi}^{-1}\},$$

which leads us to (3.9.11)

MLE

The MLE for $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{L}}$ are obtained by minimizing over the Grassmann manifolds $\mathcal{G}_{d_X,p}$ and $\mathcal{G}_{d_Y,r}$, $(\widehat{\mathbf{R}},\widehat{\mathbf{L}}) = \arg\min_{(\mathbf{R},\mathbf{L})}\{\mathrm{F}(\mathbf{R},\mathbf{L})\}$,where the objective function $\mathrm{F}(\mathbf{R},\mathbf{L})$ is

$$\begin{aligned}\mathrm{F}(\mathbf{R},\mathbf{L}) &= \log|\mathbf{R}^T\mathbf{S_X}\mathbf{R}| + \log|\mathbf{R}_0^T\mathbf{S_X}\mathbf{R}_0| + \log|\mathbf{L}_0^T\mathbf{S_Y}\mathbf{L}_0| \\ &\quad +\log|\mathbf{L}^T\left(\mathbf{S_Y} - \mathbf{S_{XY}^T}\mathbf{R}(\mathbf{R}^T\mathbf{S_X}\mathbf{R})^{-1}\mathbf{R}^T\mathbf{S_{XY}}\right)\mathbf{L}| \\ &= \log\left|\begin{array}{cc}\mathbf{R}^T\mathbf{S_X}\mathbf{R} & \mathbf{R}^T\mathbf{S_{XY}}\mathbf{L} \\ \mathbf{L}^T\mathbf{S_{XY}^T}\mathbf{R} & \mathbf{L}^T\mathbf{S_Y}\mathbf{L}\end{array}\right| + \log|\mathbf{R}_0^T\mathbf{S_X}\mathbf{R}_0| + \log|\mathbf{L}_0^T\mathbf{S_Y}\mathbf{L}_0| \\ &= \log\left|\left(\mathbf{R}^T\oplus\mathbf{L}^T\right)\begin{pmatrix}\mathbf{S_X} & \mathbf{S_{XY}} \\ \mathbf{S_{XY}^T} & \mathbf{S_Y}\end{pmatrix}(\mathbf{R}\oplus\mathbf{L})\right| \\ &\quad +\log\left|\left(\mathbf{R}^T\oplus\mathbf{L}^T\right)\left(\mathbf{S_X^{-1}}\oplus\mathbf{S_Y^{-1}}\right)(\mathbf{R}\oplus\mathbf{L})\right| \equiv \mathrm{F}(\mathbf{R}\oplus\mathbf{L}).\end{aligned}$$

Then the MLEs for $\boldsymbol{\eta},\mathbf{\Omega},\mathbf{\Omega}_0,\mathbf{\Phi},\mathbf{\Phi}_0$ are as given in the lemma.

3.9.5 Proposition 3.2

The Jacobian matrix $\mathbf{J_h}$ can be obtained by following Cook et al. (2010), and we make use of the following lemma from Cook et al. (2013).

Lemma 3.5. *Suppose that* $\mathbf{A}\in\mathbb{S}^{t\times t}$ *is non-singular and that the column-partition* $(\mathbf{B},\mathbf{B}_0)\in\mathbb{R}^{t\times t}$ *is orthogonal. Then (1)* $|\mathbf{B}^T\mathbf{A}\mathbf{B}| = |\mathbf{A}|\cdot|\mathbf{B}_0^T\mathbf{A}^{-1}\mathbf{B}_0|$*, and (2)* $|\mathbf{A}| \le |\mathbf{B}^T\mathbf{A}\mathbf{B}|\cdot|\mathbf{B}_0^T\mathbf{A}\mathbf{B}_0|$ *with equality if and only if* $\mathrm{span}(\mathbf{B})$ *reduces* $\mathbf{A}$*.*

In the population, the objective function $\mathrm{F}(\mathbf{R}\oplus\mathbf{L})$ from (3.3.4) can be written as,

$$\begin{aligned}
\mathrm{F}(\mathbf{R}\oplus\mathbf{L}) &= \log|\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}| + \log|\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R}_0| + \log|\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{L}_0| \\
&\quad + \log|\mathbf{L}^T\left(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{XY}}\right)\mathbf{L}|. \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{L}_0| + \log|\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}\mathbf{L}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}\mathbf{L}_0| + \log|\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}\mathbf{L}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}|,
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{R}(\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{R})^{-1}\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{XY}}$. The first and third inequalities came directly from Lemma 3.5, which also says that the first inequality turns to equality if and only if (C1) $\mathrm{span}(\mathbf{R})$ reduces $\boldsymbol{\Sigma}_{\mathbf{X}}$ and that the third inequality becomes equality if and only if (C3) $\mathrm{span}(\mathbf{L})$ reduces $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}$. The second and the last inequalities are obtained from the fact that $\boldsymbol{\Sigma}_{\mathbf{Y}} \geq \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}} \geq \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ for any $\mathbf{R}$. Hence, these two inequalities become equalities if and only if (C2) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{R}^T\mathbf{X}}) \subseteq \mathrm{span}(\mathbf{L})$ and (C4) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{XY}}^T) \subseteq \mathrm{span}(\mathbf{R})$. Since d_X is the dimension of the smallest subspace satisfying condition (C1) and (C4), $\mathrm{span}(\mathbf{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$. Therefore, condition (C2) and (C3) will be equivalent to (C2′) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) = \mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{XY}}) \subseteq \mathrm{span}(\mathbf{L})$, and (C3′) $\mathrm{span}(\mathbf{L})$ reduces $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. Again, because d_Y is the dimension of the smallest subspace satisfying (C2′) and (C3′), we can conclude that $\mathrm{span}(\mathbf{L}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$.

The rest of the proof relies on Shapiro's (1986) results on the asymptotics of overparameterized structural models. In order to apply Shapiro's (1986) theory in our context, we define the minimum discrepancy function as

$$\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h}) = \log|\boldsymbol{\Sigma}_{\mathbf{C}}| + \mathrm{trace}(\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\mathbf{S}_{\mathbf{C}}) - \log|\mathbf{S}_{\mathbf{C}}| - \mathrm{trace}(\mathbf{S}_{\mathbf{C}}^{-1}\mathbf{S}_{\mathbf{C}}). \tag{3.9.13}$$

Then $\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h})$ satisfies: (1) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h}) \geq 0$ for all $\widehat{\boldsymbol{h}}_{\mathrm{full}}$ and $\boldsymbol{h}$; (2) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h}) = 0$ if and only if $\widehat{\boldsymbol{h}}_{\mathrm{full}} = \boldsymbol{h}$; and (3) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h})$ is twice continuously differentiable in $\widehat{\boldsymbol{h}}_{\mathrm{full}}$ and $\boldsymbol{h}$. Recall from Section 3.4.1 that we use the subscript 0 to emphasize the true parameter: $\boldsymbol{h}_0$ and $\boldsymbol{\phi}_0$ correspond to the true distribution of $\mathbf{C}$. Then $\widehat{\boldsymbol{h}}_{\mathrm{full}}$ is $\sqrt{n}$-consistent for $\boldsymbol{h}_0$. Notice that $\widehat{\boldsymbol{h}}_{\mathrm{full}}$ is a smooth function of the sample covariance matrices which converges in distribution to the population covariance matrices, then by the delta method we know $\sqrt{n}(\widehat{\boldsymbol{h}}_{\mathrm{full}} - \boldsymbol{h}_0) \to N(0, \boldsymbol{\Gamma})$, for some positive definite covariance $\boldsymbol{\Gamma}$. Using Shapiro's (1986) Proposition 3.1 and Proposition 4.1, we will have $\sqrt{n}$-consistency results for $\mathrm{vec}(\widehat{\boldsymbol{h}})$. We next need to compare $\boldsymbol{\Gamma} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{full}})$ with $\mathbf{W} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}})$.

By noticing that $\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}} = \mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}^{1/2}$, we have

$$\begin{aligned}
\boldsymbol{\Gamma} - \mathbf{W} &= \boldsymbol{\Gamma} - \mathbf{J}_{\boldsymbol{h}}^{-1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{-1/2} \\
&= \mathbf{J}_{\boldsymbol{h}}^{-1/2}\left(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2} - \mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\right)\mathbf{J}_{\boldsymbol{h}}^{-1/2}.
\end{aligned}$$

Then the conclusion follows from the above expression.

3.9.6 Proposition 3.3

Proposition 3.3 can be seen from Proposition 3.2 by letting $\mathbf{\Gamma} = \mathbf{J}_{\boldsymbol{h}}^{-1}$ under the normality assumption.

3.9.7 Proposition 3.4

The gradient matrix

In the following computation, we switch the notation as $\boldsymbol{\beta} \to \boldsymbol{\beta}^T$ and $\boldsymbol{\eta} \to \boldsymbol{\eta}$. Then at the final result we only need to switch the matrices of Kronecker product $\mathbf{A} \otimes \mathbf{B} \to \mathbf{B} \otimes \mathbf{A}$ where $\mathbf{A} \in \mathbb{R}^{r \times r}$ and $\mathbf{B} \in \mathbb{R}^{p}$. The reason to do this is to keep consistency with the proof in Cook and Zhang (2014) for this proposition. The first three columns of the gradient matrix $\mathbf{\Delta} \equiv (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_7)$ are

$$(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_3) = \begin{pmatrix} \mathbf{L} \otimes \mathbf{R} & \mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{I}_p & \mathbf{K}_{r,p}\left(\mathbf{R}\boldsymbol{\eta} \otimes \mathbf{I}_r\right) \\ 0 & 2\mathbf{C}_p\left(\mathbf{R}\mathbf{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\mathbf{\Omega}_0\mathbf{R}_0^T\right) & 0 \\ 0 & 0 & 2\mathbf{C}_r\left(\mathbf{L}\mathbf{\Phi} \otimes \mathbf{I}_r - \mathbf{L} \otimes \mathbf{L}_0\mathbf{\Phi}_0\mathbf{L}_0^T\right) \end{pmatrix},$$

and the last four columns are

$$(\boldsymbol{\delta}_4, \ldots, \boldsymbol{\delta}_7) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \mathbf{C}_p(\mathbf{R} \otimes \mathbf{R})\mathbf{E}_{d_X} & \mathbf{C}_p(\mathbf{R}_0 \otimes \mathbf{R}_0)\mathbf{E}_{p-d_X} & 0 & 0 \\ 0 & 0 & \mathbf{C}_r(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_{d_Y} & \mathbf{C}_r(\mathbf{L}_0 \otimes \mathbf{L}_0)\mathbf{E}_{r-d_Y} \end{pmatrix},$$

where $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2 \times p^2}$ is called the contraction matrix and $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ is called the extraction matrix (Henderson and Searle, 1979). They relate the vec operation and vech operation for any symmetric matrix $\mathbf{M} \in \mathbb{S}^{p \times p}$ as: $\text{vech}(\mathbf{M}) = \mathbf{C}_p\text{vec}(\mathbf{M})$ and $\text{vec}(\mathbf{M}) = \mathbf{E}_p\text{vech}(\mathbf{M})$. More properties about the contraction and extraction matrices can be found from Henderson adn Searle (1979).

Proof. We can write $\mathbf{\Delta} = \frac{\partial \boldsymbol{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$ as $\begin{pmatrix} \partial \boldsymbol{h}_1/\partial \boldsymbol{\phi} \\ \partial \boldsymbol{h}_2/\partial \boldsymbol{\phi} \\ \partial \boldsymbol{h}_3/\partial \boldsymbol{\phi} \end{pmatrix}$ and compute it by rows.

(1) Computation of $\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{\phi}}$.

First, $\boldsymbol{h}_1(\boldsymbol{\phi}) = \text{vec}(\boldsymbol{\beta}) = \text{vec}(\mathbf{R}\boldsymbol{\eta}\mathbf{L}^T)$ and thus, $\partial \boldsymbol{h}_1(\boldsymbol{\phi})/\partial \boldsymbol{\phi}_k = 0$, $k = 4, 5, 6, 7$. The rest of the terms are:

$$\begin{aligned} \frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_1} &= \frac{\partial\left[\text{vec}(\mathbf{R}\boldsymbol{\eta}\mathbf{L}^T)\right]}{\partial \text{vec}(\boldsymbol{\eta})} = \frac{\partial\left[\{\mathbf{L} \otimes \mathbf{R}\}\text{vec}(\boldsymbol{\eta})\right]}{\partial \text{vec}(\boldsymbol{\eta})} = \mathbf{L} \otimes \mathbf{R} \\ \frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_2} &= \frac{\partial\left[\text{vec}(\mathbf{R}\boldsymbol{\eta}\mathbf{L}^T)\right]}{\partial \text{vec}(\mathbf{R})} = \frac{\partial\left[\{\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{I}_p\}\text{vec}(\mathbf{R})\right]}{\partial \text{vec}(\mathbf{R})} = \mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{I}_p \\ \frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_3} &= \frac{\partial\left[\text{vec}(\mathbf{R}\boldsymbol{\eta}\mathbf{L}^T)\right]}{\partial \text{vec}(\mathbf{L})} = \frac{\partial\left[\{\mathbf{I}_r \otimes \mathbf{R}\boldsymbol{\eta}\}\text{vec}(\mathbf{L}^T)\right]}{\partial \text{vec}(\mathbf{L})} = \{\mathbf{I}_r \otimes \mathbf{R}\mathbf{\Omega}^{-1}\boldsymbol{\eta}\}\mathbf{K}_{r,m} \\ &= \mathbf{K}_{r,p}\{\mathbf{R}\boldsymbol{\eta} \otimes \mathbf{I}_r\}. \end{aligned}$$

(2) Computation of $\partial \boldsymbol{h}_2/\partial \boldsymbol{\phi}$.

Since $\boldsymbol{h}_2(\boldsymbol{\phi}) = \text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}) = \text{vech}(\mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T)$, it follows that $\partial \boldsymbol{h}_2(\boldsymbol{\phi})/\partial \boldsymbol{\phi}_k = 0$, $k = 1, 3, 6, 7$. The rest terms can be found in Cook, Li and Chiaromonte (2010). In our notation, they are

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_2} &= 2\mathbf{C}_p(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T), \\
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_4} &= \mathbf{C}_p(\mathbf{R} \otimes \mathbf{R})\mathbf{E}_{d_X}, \\
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_5} &= \mathbf{C}_p(\mathbf{R}_0 \otimes \mathbf{R}_0)\mathbf{E}_{p-d_X}.
\end{aligned}$$

(3) Computation of $\partial \boldsymbol{h}_3/\partial \boldsymbol{\phi}$.

Since $\boldsymbol{h}_3(\boldsymbol{\phi}) = \text{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) = \text{vech}(\mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T)$, we have

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_k} &= 0, \ \text{k} = 1, 2, 4, 5, \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_3} &= 2\mathbf{C}_r(\mathbf{L}\boldsymbol{\Phi} \otimes \mathbf{I}_r - \mathbf{L} \otimes \mathbf{L}_0\boldsymbol{\Phi}_0\mathbf{L}_0^T), \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_6} &= \mathbf{C}_r(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_{d_Y}, \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_7} &= \mathbf{C}_r(\mathbf{L}_0 \otimes \mathbf{L}_0)\mathbf{E}_{r-d_Y}.
\end{aligned}$$

□

Transformation

We want to find a transformation $\boldsymbol{\Delta} = \tilde{\boldsymbol{\Delta}}\mathbf{L}$ such that $\text{span}(\boldsymbol{\Delta}) = \text{span}(\tilde{\boldsymbol{\Delta}})$ and $\tilde{\boldsymbol{\Delta}}^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}$ is block diagonal. Partition $\boldsymbol{\Delta}$ into its first three columns and its last four column, and denote them as $\boldsymbol{\Delta} \equiv (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$. It is easy to check that $\boldsymbol{\Delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}_2$ is already block diagonal. So we need to find a transformed $\tilde{\boldsymbol{\Delta}}_1$ such that $\tilde{\boldsymbol{\Delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}_1$ is block diagonal and $\tilde{\boldsymbol{\Delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}_2 = 0$.

By direct computation, we have

$$\begin{aligned}
\tilde{\boldsymbol{\Delta}}_1 &= \begin{pmatrix} \mathbf{L} \otimes \mathbf{R} & \mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0 & \mathbf{K}_{r,p}\,(\mathbf{R}\boldsymbol{\eta} \otimes \mathbf{L}_0) \\ 0 & 2\mathbf{C}_p\,(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{R}_0 - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0) & 0 \\ 0 & 0 & 2\mathbf{C}_r\,(\mathbf{L}\boldsymbol{\Phi} \otimes \mathbf{L}_0 - \mathbf{L} \otimes \mathbf{L}_0\boldsymbol{\Phi}_0) \end{pmatrix}, \\
&\equiv (\tilde{\boldsymbol{\delta}}_1, \tilde{\boldsymbol{\delta}}_2, \tilde{\boldsymbol{\delta}}_3) \\
\tilde{\boldsymbol{\Delta}}_2 &= \boldsymbol{\Delta}_2.
\end{aligned}$$

And the 7×7 blocks transformation matrix $\mathbf{L}$ can be written as a 7×3 blocks matrix $\mathbf{L}_1$ and a 7×4 block matrix $\mathbf{L}_2 = \begin{pmatrix} \mathbf{0}_{3\times 4} \\ \mathbf{I}_{4\times 4} \end{pmatrix}$ whose blocks conform to those of $\tilde{\boldsymbol{\Delta}}$. Also,

$\tilde{\boldsymbol{\Delta}}\mathbf{L}_1 = \boldsymbol{\Delta}_1$ and $\mathbf{L}_1$ is given by

$$\mathbf{L}_1 = \begin{pmatrix} \mathbf{I}_{d_X d_Y} & \boldsymbol{\eta}^T \otimes \mathbf{R}^T & \mathbf{K}_{d_Y,d_X}\left(\boldsymbol{\eta} \otimes \mathbf{L}^T\right) \\ 0 & \mathbf{I}_{d_X} \otimes \mathbf{R}_0^T & 0 \\ 0 & 0 & \mathbf{I}_{d_Y} \otimes \mathbf{L}_0^T \\ 0 & 2\mathbf{C}_{d_X}(\boldsymbol{\Omega} \otimes \mathbf{R}^T) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_{d_Y}(\boldsymbol{\Phi} \otimes \mathbf{L}^T) \\ 0 & 0 & 0 \end{pmatrix}.$$

It can be easily noticed that $\mathbf{L}$ has full row rank, hence, $\text{span}(\boldsymbol{\Delta}) = \text{span}(\tilde{\boldsymbol{\Delta}})$. Therefore, the asymptotic covariance for $\widehat{\boldsymbol{h}}$ under the envelope model is given by

$$\begin{aligned} \text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) &= \boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}\right)^{\dagger}\boldsymbol{\Delta}^T = \tilde{\boldsymbol{\Delta}}\left(\tilde{\boldsymbol{\Delta}}^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}\right)^{\dagger}\tilde{\boldsymbol{\Delta}}^T \\ &= \sum_{i=1}^{3}\tilde{\boldsymbol{\delta}}_i(\tilde{\boldsymbol{\delta}}_i^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_i)^{\dagger}\tilde{\boldsymbol{\delta}}_i^T + \sum_{j=4}^{7}\boldsymbol{\delta}_j(\boldsymbol{\delta}_j^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_j)^{\dagger}\boldsymbol{\delta}_j^T. \end{aligned}$$

Asymptotic covariance matrix

We compute the asymptotic covariance matrix for $\text{vec}(\widehat{\boldsymbol{\beta}})$. The asymptotic covariance matrix for $\text{vec}(\widehat{\boldsymbol{\beta}})$ is the upper left block of the full covariance matrix for $\widehat{\boldsymbol{h}}$. Under the standard model, it is just $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$. From Proposition 3.1, we recognize this as the asymptotic variance of $\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})$.

$$\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})) = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}. \tag{3.9.14}$$

For the envelope model, only $\tilde{\boldsymbol{\delta}}_i$, $i = 1, 2, 3$, have contributions to this block. We use the subscript $\boldsymbol{\beta\beta}$ to denote the upper left block of dimension $pr \times pr$ for a matrix and compute the corresponding three terms.

(1) Contribution from $\tilde{\boldsymbol{\delta}}_1$.

$$\begin{aligned} [\tilde{\boldsymbol{\delta}}_1(\tilde{\boldsymbol{\delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_1)^{\dagger}\tilde{\boldsymbol{\delta}}_1^T]_{\boldsymbol{\beta\beta}} &= (\mathbf{L} \otimes \mathbf{R})\left((\mathbf{L} \otimes \mathbf{R})^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}})(\mathbf{L} \otimes \mathbf{R})\right)^{\dagger}(\mathbf{L} \otimes \mathbf{R})^T \\ &= (\mathbf{L} \otimes \mathbf{R})\left(\boldsymbol{\Phi}^{-1} \otimes \boldsymbol{\Omega}\right)^{\dagger}(\mathbf{L} \otimes \mathbf{R})^T \\ &= \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T \otimes \mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T = \text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})). \end{aligned}$$

(2) Contribution from $\tilde{\boldsymbol{\delta}}_2$. Let $\mathbf{A}_1 = \mathbf{E}_p\left(\mathbf{C}_p\left(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{R}_0 - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\right)\right)$ then,

$$\begin{aligned} \tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2 &= \left(\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0\right)^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}})\left(\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0\right) + 2\mathbf{A}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{A}_1 \\ &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0 + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_{d_X} \otimes \mathbf{I}_{p-d_X} \\ &\equiv \mathbf{M}_2. \end{aligned} \tag{3.9.15}$$

Thus for the covariance,

$$[\tilde{\boldsymbol{\delta}}_2(\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2)^{\dagger}\tilde{\boldsymbol{\delta}}_2^T]_{\boldsymbol{\beta\beta}} = (\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0)(\mathbf{M}_2)^{\dagger}(\boldsymbol{\eta}\mathbf{L}^T \otimes \mathbf{R}_0^T). \tag{3.9.16}$$

(3) Contribution from $\tilde{\boldsymbol{\delta}}_3$. Let $\mathbf{A}_2 = \mathbf{E}_r\mathbf{C}_r\left(\mathbf{L}\boldsymbol{\Phi}\otimes\mathbf{L}_0 - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Phi}_0\right)$ then,

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3 &= \left(\mathbf{K}_{r,p}\left(\mathbf{R}\boldsymbol{\eta}\otimes\mathbf{L}_0\right)\right)^T\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{Y}|\mathbf{X}}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}\right)\left(\mathbf{K}_{r,p}\left(\mathbf{R}\boldsymbol{\eta}\otimes\mathbf{L}_0\right)\right)\\
&+ 2\mathbf{A}^T(\boldsymbol{\Sigma}^{-1}_{\mathbf{Y}|\mathbf{X}}\otimes\boldsymbol{\Sigma}^{-1}_{\mathbf{Y}|\mathbf{X}})\mathbf{A}\\
&= \boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta}\otimes\boldsymbol{\Phi}_0^{-1} + \boldsymbol{\Phi}\otimes\boldsymbol{\Phi}_0^{-1} + \boldsymbol{\Phi}^{-1}\otimes\boldsymbol{\Phi}_0 - 2\mathbf{I}_{d_Y}\otimes\mathbf{I}_{r-d_Y}\\
&\equiv \mathbf{M}_3.
\end{aligned}$$

Thus for the covariance,

$$\begin{aligned}
[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3)^\dagger\tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta\beta}} &= \mathbf{K}_{r,p}\left(\mathbf{R}\boldsymbol{\eta}\otimes\mathbf{L}_0\right)\left(\mathbf{M}_3\right)^\dagger\left(\mathbf{K}_{r,p}\left(\mathbf{R}\boldsymbol{\eta}\otimes\mathbf{L}_0\right)\right)^T\\
&= \left(\mathbf{L}_0\otimes\mathbf{R}\boldsymbol{\eta}\right)\mathbf{K}_{r-d_Y,d_Y}\left(\mathbf{M}_3\right)^\dagger\mathbf{K}^T_{r-d_Y,d_Y}\left(\boldsymbol{\eta}^T\mathbf{R}^T\otimes\mathbf{L}_0^T\right)\\
&= \left(\mathbf{L}_0\otimes\mathbf{R}\boldsymbol{\eta}\right)\left(\mathbf{K}^T_{r-d_Y,d_Y}\mathbf{M}_3\mathbf{K}_{r-d_Y,d_Y}\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{R}^T\otimes\mathbf{L}_0^T\right)
\end{aligned}$$

the last equality is because $\mathbf{K}_{r-d_Y,d_Y}$ is non-singular and $\mathbf{K}_{r-d_Y,d_Y}\mathbf{K}^T_{r-d_Y,d_Y} = \mathbf{I}_{(r-d_Y)d_Y}$. Therefore,

$$\begin{aligned}
[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3)^\dagger\tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta\beta}} &= \left(\mathbf{L}_0\otimes\mathbf{R}\boldsymbol{\eta}\right)\left(\mathbf{K}^T_{r-d_Y,d_Y}\mathbf{M}_3\mathbf{K}_{r-d_Y,d_Y}\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{R}^T\otimes\mathbf{L}_0^T\right)\\
&= \left(\mathbf{L}_0\otimes\mathbf{R}\boldsymbol{\eta}\right)\left(\mathbf{M}_4\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{R}^T\otimes\mathbf{L}_0^T\right),
\end{aligned}$$

where

$$\mathbf{M}_4 = \boldsymbol{\Phi}_0^{-1}\otimes\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta} + \boldsymbol{\Phi}_0^{-1}\otimes\boldsymbol{\Phi} + \boldsymbol{\Phi}_0\otimes\boldsymbol{\Phi}^{-1} - 2\mathbf{I}_{r-d_Y}\otimes\mathbf{I}_{d_Y}. \tag{3.9.17}$$

Interpretations

The Fisher information matrix for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T,\ldots,\boldsymbol{\phi}_7)^T$, $\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}$, has the form

$$\begin{pmatrix}
\mathbf{J}_{\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\eta}\mathbf{R}} & \mathbf{J}_{\boldsymbol{\eta}\mathbf{L}} & 0 & 0 & 0 & 0\\
\mathbf{J}^T_{\boldsymbol{\eta}\mathbf{R}} & \mathbf{J}_{\mathbf{R}} & \mathbf{J}_{\mathbf{R}\mathbf{L}} & \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}} & 0 & 0 & 0\\
\mathbf{J}^T_{\boldsymbol{\eta}\mathbf{L}} & \mathbf{J}^T_{\mathbf{R}\mathbf{L}} & \mathbf{J}_{\mathbf{L}} & 0 & 0 & \mathbf{J}_{\mathbf{L}\boldsymbol{\Phi}} & 0\\
0 & \mathbf{J}^T_{\mathbf{R}\boldsymbol{\Omega}} & 0 & \mathbf{J}_{\boldsymbol{\Omega}} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & \mathbf{J}_{\boldsymbol{\Omega}_0} & 0 & 0\\
0 & 0 & \mathbf{J}^T_{\mathbf{L}\boldsymbol{\Phi}} & 0 & 0 & \mathbf{J}_{\boldsymbol{\Phi}} & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \mathbf{J}_{\boldsymbol{\Phi}_0}
\end{pmatrix}.$$

If $\mathbf{R}$ and $\mathbf{L}$ are known, then the asymptotic variance of the MLE of vec($\boldsymbol{\eta}$), denoted by vec($\widehat{\boldsymbol{\eta}}_{\mathbf{R},\mathbf{L}}$), is simply $\mathbf{J}_{\boldsymbol{\eta}}^{-1} = \boldsymbol{\Phi}\otimes\boldsymbol{\Omega}^{-1}$. As we discussed before,

$$\begin{aligned}
\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R},\mathbf{L}})) &= \text{avar}(\sqrt{n}\text{vec}(\mathbf{R}\widehat{\boldsymbol{\eta}}_{\mathbf{R},\mathbf{L}}\mathbf{L}^T))\\
&= (\mathbf{L}\otimes\mathbf{R})\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\eta}}_{\mathbf{R},\mathbf{L}}))(\mathbf{L}^T\otimes\mathbf{R}^T)\\
&= \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^T\otimes\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T.
\end{aligned}$$

If $\boldsymbol{\eta}$ and $\mathbf{L}$ are known, then the asymptotic variance of the MLE of vec($\mathbf{R}$), denoted by vec($\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}}$), is $\left(\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T\right)^{-1}$. Notice that

$$\begin{aligned}
\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T &= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2 - \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2 \\
&= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}^{1/2}\left(\mathbf{I} - \mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4}\right)\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_2 \\
&= \boldsymbol{\delta}_2^T\left(\mathbf{J}_{\boldsymbol{h}} - \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\right)\boldsymbol{\delta}_2 \\
&\equiv \boldsymbol{\delta}_2^T\left(\mathbf{J}_{\boldsymbol{h}} - \mathbf{J}_0\right)\boldsymbol{\delta}_2.
\end{aligned}$$

In the second equality above, $\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4}$ is the projection matrix onto $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4$. But it not obvious how the projection might facilitate the calculation here. We look into the last equality, recall that $\boldsymbol{\delta}_4^T = (0, (\mathbf{C}_p(\mathbf{R}\otimes\mathbf{R})\mathbf{E}_{d_X})^T, 0)$ and that $\mathbf{J}_{\boldsymbol{h}}$ was given in (3.4.1). Therefore,

$$\begin{aligned}
\mathbf{J}_0 &= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\mathbf{R}^T\otimes\mathbf{R}^T)\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{P}_{\mathbf{E}_{d_X}}(\mathbf{R}\otimes\mathbf{R})\mathbf{E}_p\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\mathbf{R}^T\otimes\mathbf{R}^T)(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})(\mathbf{R}\otimes\mathbf{R})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}.
\end{aligned}$$

The matrix $\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T$ has 3×3 blocks. Only its middle $[\cdot]_{22}$ block is nonzero, other 8 blocks are all zeros. Then we compute the middle block as following, where we let $\mathbf{A}_3 = \mathbf{C}_p(\mathbf{R}\otimes\mathbf{R})\mathbf{E}_{d_X}$.

$$\begin{aligned}
&\left[\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T\right]_{22} && (3.9.18)\\
&= \mathbf{A}_3\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\mathbf{E}_{d_X}^T\mathbf{A}_3^T \\
&= 2\mathbf{A}_3\mathbf{E}_{d_X}\left(\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\mathbf{E}_{d_X}^T\mathbf{A}_3^T \\
&= 2\mathbf{A}_3\mathbf{P}_{\mathbf{E}_{d_X}}(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})\mathbf{P}_{\mathbf{E}_{d_X}}\mathbf{A}_3^T && (3.9.19)\\
&= 2\mathbf{C}_p(\mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T\otimes\mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T)\mathbf{C}_p^T.
\end{aligned}$$

Where the equation (3.9.19) is obtained using Corollary E.1. in CLC (2010; supplementary material): $\mathbf{E}_{d_X}\in\mathbb{R}^{d_X^2\times d_X(d_X+1)/2}$ is nonsingular and $\mathbf{P}_{\mathbf{E}_l}$ commute with $(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})$, then $\mathbf{E}_{d_X}\left(\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{-1}\mathbf{E}_{d_X}^T = \mathbf{P}_{\mathbf{E}_{d_X}}\left[(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\right]^{-1}\mathbf{P}_{\mathbf{E}_{d_X}}$.

Now, $\mathbf{J}_0 = \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}$ only has its $[\cdot]_{22}$ block nonzero. For notational convenience, let $\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\mathbf{B} = \mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T\otimes\mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T$. Then

$$\begin{aligned}
[\mathbf{J}_0]_{22} &= \left[\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\right]_{22} = 2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{E}_p\times 2\mathbf{C}_p\mathbf{B}\mathbf{C}_p^T\times 2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{E}_p \\
&= 2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}\mathbf{E}_p = 2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{B}\mathbf{A}\mathbf{E}_p = 2^{-1}\mathbf{E}_p^T\mathbf{B}\mathbf{E}_p.
\end{aligned}$$

Then, recall that $\boldsymbol{\delta}_2 = \begin{pmatrix} \mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{I}_p \\ 2\mathbf{C}_p\left(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T\right) \\ 0 \end{pmatrix}$ and,

$$\begin{aligned}
\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T &= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2 - \boldsymbol{\delta}_2^T\mathbf{J}_0\boldsymbol{\delta}_2 \\
&= \left(\boldsymbol{\eta}\mathbf{L}^T \otimes \mathbf{I}_p\right)\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}\right)\left(\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{I}_p\right) \\
&\quad +2\left(\boldsymbol{\Omega}\mathbf{R}^T \otimes \mathbf{I}_p - \mathbf{R}^T \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T\right)\mathbf{P}_{\mathbf{E}_p}\{(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}) - \\
&\quad (\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T)\}\mathbf{P}_{\mathbf{E}_p}\left(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T\right) \\
&= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}_{\mathbf{X}} + 2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}.
\end{aligned}$$

The second term $2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}$ is defined by $\mathbf{A} = \mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T$, and

$$\begin{aligned}
\mathbf{B} &= (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}) - (\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T) \\
&= \mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T + \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T.
\end{aligned}$$

Then we compute this term directly by noticing $\mathbf{P}_{\mathbf{E}_p} = \frac{1}{2}(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})$ and $\mathbf{B}$ commutes with $\mathbf{P}_{\mathbf{E}_p}$, $2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A} = 2\mathbf{A}^T\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A} = \mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})\mathbf{A}$. Then, notice that in $\mathbf{A}^T\mathbf{B}$ only the first term in $\mathbf{B}$ will not vanish. Hence,

$$\begin{aligned}
\mathbf{A}^T\mathbf{B} &= (\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T)^T\mathbf{R}\boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T \\
&= \mathbf{R}^T \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T - \boldsymbol{\Omega}^{-1}\mathbf{R}^T \otimes \mathbf{R}_0\mathbf{R}_0^T.
\end{aligned}$$

And,

$$\begin{aligned}
\mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})\mathbf{A} &= \mathbf{A}^T\mathbf{B}\left(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T\right) \\
&\quad +\mathbf{A}^T\mathbf{B}\left(\mathbf{I}_p \otimes \mathbf{R}\boldsymbol{\Omega} - \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T \otimes \mathbf{R}\right)\mathbf{K}_{l,p} \\
&= \mathbf{A}^T\mathbf{B}\left(\mathbf{R}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T\right) \\
&= \boldsymbol{\Omega} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T - \mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T - \mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T + \boldsymbol{\Omega}^{-1} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T \\
&= \boldsymbol{\Omega} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T - 2\mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T + \boldsymbol{\Omega}^{-1} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T.
\end{aligned}$$

So far, we have found,

$$\begin{aligned}
\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}_{\mathbf{X}} + 2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A} \\
&= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}_{\mathbf{X}} + \mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2} + \mathbf{K}_{p,p})\mathbf{A} \\
&= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}_{\mathbf{X}} + \boldsymbol{\Omega} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0^{-1}\mathbf{R}_0^T - 2\mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T + \boldsymbol{\Omega}^{-1} \otimes \mathbf{R}_0\boldsymbol{\Omega}_0\mathbf{R}_0^T.
\end{aligned}$$

Comparing this expression with (3.9.15) and (3.9.16), we have

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2 &= (\mathbf{I}_{d_X} \otimes \mathbf{R}_0^T)\left(\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T\right)(\mathbf{I}_{d_X} \otimes \mathbf{R}_0) \\
&= (\mathbf{I}_{d_X} \otimes \mathbf{R}_0^T)\left(\text{avar}(\sqrt{n}\text{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}}))\right)^{\dagger}(\mathbf{I}_{d_X} \otimes \mathbf{R}_0).
\end{aligned}$$

Consequently,

$$\begin{aligned}
& [\tilde{\boldsymbol{\delta}}_2(\tilde{\boldsymbol{\delta}}_2^T \mathbf{J}_{\boldsymbol{h}} \tilde{\boldsymbol{\delta}}_2)^\dagger \tilde{\boldsymbol{\delta}}_2^T]_{\boldsymbol{\beta\beta}} \\
= \; & (\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0) \left((\mathbf{I}_{d_X} \otimes \mathbf{R}_0^T) \left(\text{avar}(\sqrt{n}\text{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}})) \right)^\dagger (\mathbf{I}_{d_X} \otimes \mathbf{R}_0) \right)^\dagger (\boldsymbol{\eta}\mathbf{L}^T \otimes \mathbf{R}_0^T) \\
= \; & (\mathbf{L}\boldsymbol{\eta}^T \otimes \mathbf{R}_0\mathbf{R}_0^T) \left(\text{avar}(\sqrt{n}\text{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}})) \right) (\boldsymbol{\eta}\mathbf{L}^T \otimes \mathbf{R}_0\mathbf{R}_0^T) \qquad (3.9.20) \\
= \; & \text{avar}\left(\sqrt{n}\text{vec}(\mathbf{Q}_{\mathbf{R}} \widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}} \boldsymbol{\eta}\mathbf{L}^T) \right) = \text{avar}\left(\sqrt{n}\text{vec}(\mathbf{Q}_{\mathbf{R}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}) \right).
\end{aligned}$$

where (3.9.20) can be obtained using Corollary E.1. in CLC (2010; supplementary material): $\mathbf{I}_{d_X} \otimes \mathbf{R}_0$ has full column rank, and its projection $\mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T$ commute with $\mathbf{C} \equiv \left(\text{avar}(\sqrt{n}\text{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}})) \right)^\dagger = \left(\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}} \mathbf{J}_{\boldsymbol{\Omega}}^{-1} \mathbf{J}_{\mathbf{R}\boldsymbol{\Omega}}^T \right)$, then

$$\mathbf{I}_{d_X} \otimes \mathbf{R}_0 \left(\left(\mathbf{I}_{d_X} \otimes \mathbf{R}_0^T \right) \mathbf{C} \left(\mathbf{I}_{d_X} \otimes \mathbf{R}_0 \right) \right)^{-1} \mathbf{I}_{d_X} \otimes \mathbf{R}_0^T = \left(\mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T \right) \mathbf{C}^{-1} \left(\mathbf{I}_{d_X} \otimes \mathbf{R}_0\mathbf{R}_0^T \right).$$

If $\boldsymbol{\eta}$ and $\mathbf{R}$ are known, then the asymptotic variance of the MLE of vec($\mathbf{L}$), denoted by vec($\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}$), is $\left(\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Phi}} \mathbf{J}_{\boldsymbol{\Phi}}^{-1} \mathbf{J}_{\mathbf{L}\boldsymbol{\Phi}}^T \right)^{-1}$. Either by similar computation or by symmetry, we have

$$\begin{aligned}
[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T \mathbf{J}_{\boldsymbol{h}} \tilde{\boldsymbol{\delta}}_3)^\dagger \tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta\beta}} &= \text{avar}\left(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}} \mathbf{Q}_{\mathbf{L}}) \right) = \text{avar}\left(\sqrt{n}\text{vec}(\mathbf{R}\boldsymbol{\eta}\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}^T \mathbf{Q}_{\mathbf{L}}) \right) \\
&= (\mathbf{L}_0\mathbf{L}_0^T \otimes \mathbf{R}\boldsymbol{\eta}) \text{avar}\left(\sqrt{n}\text{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}^T) \right) (\mathbf{L}_0\mathbf{L}_0^T \otimes \boldsymbol{\eta}^T\mathbf{R}^T).
\end{aligned}$$

Chapter 4

Envelopes and Reduced-rank Regression

4.1 Introduction

The multivariate linear regression model for $p \times 1$ non-stochastic predictor $\mathbf{X}$ and $r \times 1$ stochastic response $\mathbf{Y}$ can be written as

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{4.1.1}$$

where the error vector $\boldsymbol{\epsilon}$ has mean zero and covariance matrix $\boldsymbol{\Sigma} > 0$ and is independent of $\mathbf{X}$. In this chapter, we use $\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ to denote the error covariance, to avoid notation proliferations. Our interest in this model still lies in prediction and in studying the interrelation between $\mathbf{X}$ and $\mathbf{Y}$ through the regression coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$. There is a general awareness that the estimation of $\boldsymbol{\beta}$ may often be improved by reducing the dimensionalities of $\mathbf{X}$ and $\mathbf{Y}$, and reduced-rank regression is popular method for doing so. We propose a reduced-rank envelope model that extends the nascent idea of envelopes to reduced-rank regression. The purpose of this chapter is to integrate reduced-rank regression and envelopes, resulting in an overarching method that can choose the better of the two methods when appropriate and that has the potential to perform better than either of them.

Reduced-rank regression (Anderson 1951; Izenman 1975; Reinsel and Velu 1998) arises frequently in multivariate statistical analysis, and has been applied widely across the applied sciences. By restricting the rank of the regression coefficient matrix $\text{rank}(\boldsymbol{\beta}) = d < \min(r, p)$, the total number of parameters is reduced and efficiency in estimation is improved. The analysis of reduced-rank regression (Izenman 1975; Tso 1981; Reinsel and Velu 1998; Anderson 2002) connects with many important multivariate methods such as principal components analysis, canonical correlation analysis and multiple time series modeling. The asymptotic advantages of the reduced-rank regression estimator over the

standard ordinary least squares estimator were studied by Stoica and Viberg (1996) and Anderson (1999). Chen et al. (2012) and Chen and Huang (2012) extended reduced-rank regression to high-dimensional settings and demonstrated the advantages of parsimoniously reducing model parameters and interrelating response variables.

Envelope regression, which was first proposed by Cook et al. (2010), is another way of parsimoniously reducing the total number of parameters from the standard model (4.1.1) and gaining both efficiency in estimation and accuracy in prediction. The key idea of envelopes is to identify and eliminate information in the responses and the predictors that is immaterial to the estimation of $\boldsymbol{\beta}$ but still introduces unnecessary variation into estimation. Envelope reduction can be effective even when $d = \min(p, r)$, which is the case where reduced-rank regression has no reduction.

Envelope and reduced-rank regressions have different perspectives on dimension reduction. It may take considerable effort to find which method is more efficient for a problem in practice. The proposed reduced-rank envelope model combines the strengths of envelopes and reduced-rank regression, which mitigates the burden of selecting among the two methods. When one of the two methods behaves poorly, the reduced-rank envelope model automatically degenerates towards the other one; when both methods show efficiency gains, the reduced-rank envelope estimator will enjoy a synergy from combining the two approaches and may improve over both estimators.

The rest of this chapter is organized as follows. In Section 4.2, we review and summarize some fundamental results for reduced-rank regression and envelopes that are relevant to our development. We set up our reduced-rank envelope model in Section 4.2.2, where we also give intuitive connections to reduced-rank regression and envelope models. In Section 4.3.1, we summarize parameterizations for each model and show that the total number of parameters in the reduced-rank envelope model is fewer than that of the other models. Likelihood-based estimators for the reduced-rank envelope model are derived in Section 4.3.2. Asymptotic properties are studied in Section 4.4. We show that the reduced-rank envelope estimator is asymptotically more efficient than ordinary least squares, reduced-rank regression and envelope estimators under normal errors, and is still $\sqrt{n}$-consistent without the normality assumption. Section 4.5 discusses procedures for selecting the rank of the coefficient matrix and the dimension of the envelope. Encouraging simulation results and real data examples are presented in Section 4.6 and 4.7. Proofs and other technical details are included in Section 4.9.

4.2 Reduced-rank envelope model

4.2.1 Reduced-rank regression

Reduced-rank regression allows that $\text{rank}(\boldsymbol{\beta}) = d < \min(p, r)$ so that we can write the model parameterization as

$$\boldsymbol{\beta} = \mathbf{AB},\ \mathbf{A} \in \mathbb{R}^{r\times d},\ \mathbf{B} \in \mathbb{R}^{d\times p},\ \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = d, \tag{4.2.1}$$

where no additional constraints are imposed on $\mathbf{A}$ or $\mathbf{B}$. The maximum likelihood estimators for the reduced-rank regression parameters were derived by Anderson (1999), Reinsel and Velu (1998) and Stoica and Viberg (1996), under various constraints on $\mathbf{A}$ and $\mathbf{B}$ for identifiability, such as $\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T = \mathbf{I}_d$ or $\mathbf{A}^T\mathbf{A} = \mathbf{I}_d$. The decomposition $\boldsymbol{\beta} = \mathbf{AB}$ is still non-unique even with those identifiable constraints: for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$, $\mathbf{A}_1 = \mathbf{AO}$ and $\mathbf{B}_1 = \mathbf{O}^T\mathbf{B}$ offer another valid decomposition that satisfies the constraints. The parameters of interests, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, are nevertheless identifiable as well as $\text{span}(\mathbf{A}) = \text{span}(\boldsymbol{\beta})$ and $\text{span}(\mathbf{B}^T) = \text{span}(\boldsymbol{\beta}^T)$. We present this article in an apparently novel unified framework so that every statement involving $\mathbf{A}$ or $\mathbf{B}$ holds universally for any decomposition $\boldsymbol{\beta} = \mathbf{AB}$ satisfying (4.2.1).

The log-likelihood of model (4.1.1) under normality of $\boldsymbol{\epsilon}$ can be written as,

$$L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \simeq -\frac{n}{2}\left\{\log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\beta}\mathbf{X}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\alpha} - \boldsymbol{\beta}\mathbf{X}_i)\right\}, \tag{4.2.2}$$

which is to be maximized under the constraint that $\text{rank}(\boldsymbol{\beta}) = d$, or equivalently under the parameterization $\boldsymbol{\beta} = \mathbf{AB}$. The symbol $\simeq$ denotes an equality from which any unimportant additive constant has been eliminated. We treat $L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ as a general purpose objective function, which will be maximized under (4.2.1). The following lemma summarizes the reduced-rank regression estimator that maximizes (4.2.2). Rigorous derivation can be found in Anderson (1999).

We define the sample canonical correlation matrix between $\mathbf{Y}$ and $\mathbf{X}$ as $\mathbf{C}_{\mathbf{YX}} = \mathbf{S}_{\mathbf{Y}}^{-1/2}\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1/2}$ and $\mathbf{C}_{\mathbf{XY}} = \mathbf{C}_{\mathbf{YX}}^T$. Truncated matrices are represented with superscripts. For example, $\mathbf{C}_{\mathbf{YX}}^{(d)}$ and $\mathbf{S}_{\mathbf{YX}}^{(d)}$ are constructed by truncated singular value decompositions of $\mathbf{C}_{\mathbf{YX}}$ and $\mathbf{S}_{\mathbf{YX}}$ with only the largest d singular values being kept.

Lemma 4.1. *Under the reduced-rank regression parameterization (4.2.1), the likelihood-based objective function from (4.2.2) is maximized at* $\widehat{\boldsymbol{\alpha}}_{\text{RR}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\text{RR}}\overline{\mathbf{X}}$ *and*

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{RR}} &= \mathbf{S}_{\mathbf{Y}}^{1/2}\mathbf{C}_{\mathbf{YX}}^{(d)}\mathbf{S}_{\mathbf{X}}^{-1/2}\\ \widehat{\boldsymbol{\Sigma}}_{\text{RR}} &= \mathbf{S}_{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\text{RR}}\mathbf{S}_{\mathbf{XY}} = \mathbf{S}_{\mathbf{Y}}^{1/2}\left\{\mathbf{I}_r - \mathbf{C}_{\mathbf{YX}}^{(d)}\mathbf{C}_{\mathbf{XY}}^{(d)}\right\}\mathbf{S}_{\mathbf{Y}}^{1/2}.\end{aligned}$$

There are a variety forms of maximizers $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ in the literature under different constraints on $\mathbf{A}$ and $\mathbf{B}$. They could all be reproduced by decomposing the rank-d estimator

$\widehat{\boldsymbol{\beta}}_{\mathrm{RR}}$ in Lemma 4.1. The ordinary least squares estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be written as $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{S}_{\mathbf{Y}}^{1/2}\mathbf{C}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1/2}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{OLS}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{Y}}^{1/2}\left\{\mathbf{I}_r - \mathbf{C}_{\mathbf{YX}}\mathbf{C}_{\mathbf{XY}}\right\}\mathbf{S}_{\mathbf{Y}}^{1/2}$ by replacing the truncated sample canonical correlation matrices $\mathbf{C}_{(\cdot)}^{(d)}$ with the untruncated ones $\mathbf{C}_{(\cdot)}$. This Lemma also reveals the scale equivariant property of both reduced-rank regression and ordinary least squares estimators since the truncated sample canonical correlation matrices are scale invariant.

4.2.2 Reduced-rank envelope model

Let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be an orthogonal basis for $\mathbb{R}^r$ so that $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$. Then $\dim(\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})) = u$ and

$$\boldsymbol{\beta} = \mathbf{AB} = \boldsymbol{\Gamma}\boldsymbol{\xi} = \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{B},\ \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{4.2.3}$$

where $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are symmetric positive definite matrices in $\mathbb{R}^{u\times u}$ and $\mathbb{R}^{(r-u)\times(r-u)}$ respectively and $\boldsymbol{\eta} \in \mathbb{R}^{u\times d}$, $u \geq d$, are the coordinates of $\mathbf{A}$ with respect to $\boldsymbol{\Gamma}$. The parameterization $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\xi}$ with $\boldsymbol{\xi} \in \mathbb{R}^{u\times p}$ occurs in the envelope model of Cook et al. (2010). We still impose no additional constraint on $\mathbf{A}$, $\mathbf{B}$ or $\boldsymbol{\eta}$ other than requiring them all to have rank d. The decompositions of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in (4.2.3) are not unique but $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unique.

To see the connections between the reduced-rank envelope model and reduced-rank regression, we next consider the situation in which $\boldsymbol{\Gamma}$ is known. Notice that $\mathrm{span}(\boldsymbol{\Gamma})$ is uniquely defined while $\boldsymbol{\Gamma}$ is unique up to an orthogonal transformation in $\mathbb{R}^u$. Although expressions in Lemma 4.2 are given in terms of $\boldsymbol{\Gamma}$, the final estimators $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$ depend on $\boldsymbol{\Gamma}$ only via $\mathrm{span}(\boldsymbol{\Gamma})$: for any orthogonal transformation $\mathbf{O} \in \mathbb{R}^{u\times u}$, we have $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}\mathbf{O}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}\mathbf{O}}$.

Lemma 4.2. *Under the reduced-rank envelope model (4.2.3), the likelihood-based objective function from (4.2.2) with given $\boldsymbol{\Gamma}$ is maximized at $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Gamma}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}\overline{\mathbf{X}}$ and*

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2}\mathbf{C}_{\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{S}_{\mathbf{X}}^{-1/2} \\
\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2}\left\{\mathbf{I}_u - \mathbf{C}_{\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{C}_{\mathbf{X},\boldsymbol{\Gamma}^T\mathbf{Y}}^{(d)}\right\}\mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2}\boldsymbol{\Gamma}^T + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\boldsymbol{\Gamma}}.
\end{aligned}$$

The implication of Lemma 4.2 is clear: once we know the envelope, we can focus our attention on the reduced response $\boldsymbol{\Gamma}^T\mathbf{Y}$ and find $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}}$, which is the rank-$d$ reduced-rank regression estimator of $\boldsymbol{\Gamma}^T\mathbf{Y}$ on $\mathbf{X}$. By Definition 1.1, the covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$ is now reduced by $\mathrm{span}(\boldsymbol{\Gamma})$ since $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}\mathbf{Q}_{\boldsymbol{\Gamma}}$. Hence $\mathrm{span}(\boldsymbol{\Gamma})$ is a reducing subspace of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Gamma}}$ that also contains $\mathrm{span}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$, and the envelope structure is preserved by the construction of these estimators. In Section 4.3.2, we derive the likelihood-based estimator $\widehat{\boldsymbol{\Gamma}}$ and demonstrate that the reduced-rank envelope estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ coincide with the estimators in Lemma 4.2 by replacing $\boldsymbol{\Gamma}$ with $\widehat{\boldsymbol{\Gamma}}$.

When the envelope dimension $u = r$, there is no immaterial information to be reduced by the envelope method. Then the reduced-rank envelope model degenerates to the reduced-rank regression (4.2.1), $\boldsymbol{\Gamma} = \mathbf{I}_r$. When the regression coefficient matrix is full rank $\text{rank}(\boldsymbol{\beta}) = p \leq r$, reduced-rank regression is equivalent to ordinary least squares and the reduced-rank envelope model degenerates to the ordinary envelope model. Two extreme situations are then: (a) if $p > r = 1$ then both methods degenerate to the standard method, which produces no reduction; (b) if $r > p = 1$ then reduced-rank regression can not provide any response reduction while reduced-rank envelopes can still gain efficiency by projecting the response onto the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. The reduced-rank envelope model can be extended to the predictor envelopes by Cook et al. (2013), so that it can resolve the problem in (a) and provide potential gain by enveloping in the predictor space.

4.3 Likelihood-based estimation for reduced-rank envelope

4.3.1 Parameters in different models

Following Cook, Li and Chiaromonte (2010), we define the following estimable functions $\mathbf{h}$ for the standard model (4.1.1), parameters $\boldsymbol{\psi}$ for the reduced-rank model, parameters $\boldsymbol{\delta}$ for the envelope model and parameter $\boldsymbol{\phi}$ for the reduced-rank envelope model. The common parameter $\boldsymbol{\alpha}$ is omitted because its estimator takes the following form for all methods: $\widehat{\boldsymbol{\alpha}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}\overline{\mathbf{X}}$, while $\overline{\mathbf{Y}}$ and $\overline{\mathbf{X}}$ are asymptotically independent of the other estimators.

$$\mathbf{h} = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \ \boldsymbol{\psi} = \begin{pmatrix} \text{vec}(\mathbf{A}) \\ \text{vec}(\mathbf{B}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \ \boldsymbol{\delta} = \begin{pmatrix} \text{vec}(\boldsymbol{\Gamma}) \\ \text{vec}(\boldsymbol{\xi}) \\ \text{vech}(\boldsymbol{\Omega}) \\ \text{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \ \boldsymbol{\phi} = \begin{pmatrix} \text{vec}(\boldsymbol{\Gamma}) \\ \text{vec}(\boldsymbol{\eta}) \\ \text{vec}(\mathbf{B}) \\ \text{vech}(\boldsymbol{\Omega}) \\ \text{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \tag{4.3.1}$$

where we define $\mathbf{h} = (\mathbf{h}_1^T, \mathbf{h}_2^T)^T$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T)^T$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_4^T)^T$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_5)^T$ correspondingly. We have $\mathbf{h} = \mathbf{h}(\boldsymbol{\psi})$ under the reduced-rank model, $\mathbf{h} = \mathbf{h}(\boldsymbol{\delta})$ under the envelope model and $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi})$ under the reduced-rank envelope model.

We use $\mathcal{N}(\cdot)$ to denote the total number of unique real parameters in a vector of model parameters. We have the following summary for each method:

(i) standard linear model, $\mathcal{N}_{\text{OLS}} := \mathcal{N}(\mathbf{h}) = pr + r(r+1)/2$;

(ii) reduced-rank model, $\mathcal{N}_{\text{RR}} := \mathcal{N}(\boldsymbol{\psi}) = (p + r - d)d + r(r+1)/2$;

(iii) envelope model, $\mathcal{N}_{\text{env}} := \mathcal{N}(\boldsymbol{\delta}) = pu + r(r+1)/2$;

(iv) reduced-rank envelope model, $\mathcal{N}_{\text{RE}} := \mathcal{N}(\boldsymbol{\phi}) = (p + u - d)d + r(r+1)/2$.

By straightforward calculation we observe that the total number of unique parameters is reduced by $(p-d)(r-d) \geq 0$ from standard model to reduced-rank regression, and is further reduced by $(r-u)d \geq 0$ from reduced-rank regressionto reduced-rank envelopes. Similarly, the total number of unique parameters is reduced by $p(r-u) \geq 0$ from the standard model to envelopes, and is further reduced by $(p-d)(u-d) \geq 0$ from the envelope model to the reduced-rank envelope model.

4.3.2 Estimators for the reduced-rank envelope model parameters

The goal of this section is to derive the reduced-rank envelope estimators for given d and u. Procedures for selecting d and u are discussed in Section 4.5. The likelihood-based reduced-rank envelope estimators is obtained by substituting $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi})$ into (4.2.2) and maximizing $L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}(\boldsymbol{\phi}), \boldsymbol{\Sigma}(\boldsymbol{\phi})) \equiv L_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma}|d, u)$ over all parameters except $\boldsymbol{\Gamma}$ because they live on a product space and the optimizing value of $\boldsymbol{\Gamma}$ cannot be found analytically. We then arrive at the estimator $\widehat{\boldsymbol{\Gamma}}$ from optimization over a Grassmannian as described in the following Proposition. For any semi-orthogonal $r \times u$ matrix $\mathbf{G}$, we define $\mathbf{Z}_\mathbf{G} = (\mathbf{G}^T\mathbf{S}_\mathbf{Y}\mathbf{G})^{-1/2}\mathbf{G}^T\mathbf{Y}$ to be the standardized version of $\mathbf{G}^T\mathbf{Y} \in \mathbb{R}^u$ with sample covariance $\mathbf{I}_u$, and let $\widehat{\omega}_i(\mathbf{G})$, $i = 1, \ldots, u$, be the i-th eigenvalue of $\mathbf{S}^{-1}_{\mathbf{Z}_\mathbf{G}|\mathbf{X}} = (\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2}(\mathbf{G}^T\mathbf{S}_\mathbf{Y}\mathbf{G})(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2}$.

Proposition 4.1. *The estimator* $\widehat{\boldsymbol{\Gamma}} = \arg\min_\mathbf{G} \mathrm{F}_n(\mathbf{G}|d, u)$ *is the maximizer of*

$$L_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma}|d, u),$$

where the optimization is over $\mathcal{G}_{r,u}$ *and*

$$\begin{aligned} \mathrm{F}_n(\mathbf{G}|d,u) &= \log|\mathbf{G}^T\mathbf{S}_\mathbf{Y}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_\mathbf{Y}^{-1}\mathbf{G}| + \log|\mathbf{I}_u - \mathbf{S}^{(d)}_{\mathbf{Z}_\mathbf{G}\circ\mathbf{X}}| && (4.3.2)\\ &= \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_\mathbf{Y}^{-1}\mathbf{G}| + \sum_{i=d+1}^{u}\log[\widehat{\omega}_i(\mathbf{G})]. && (4.3.3)\end{aligned}$$

We find in practice that the form of objective function (4.3.3) can be more easily and stably evaluated than (4.3.2). The analytical expression of $\partial\mathrm{F}_n(\mathbf{G}|d,u)/\partial\mathbf{G}$ based on (4.3.3) is used to facilitate the Newton-Raphson or conjugate gradient iterations. The formulation in (4.3.2) describes some operating characteristics of the reduced-rank envelope objective function. Lemma 4.1 and the relationship $\mathbf{S}^{(d)}_{\mathbf{Z}_\mathbf{G}\circ\mathbf{X}} = \mathbf{C}^{(d)}_{\mathbf{Z}_\mathbf{G}\mathbf{X}}\mathbf{C}^{(d)}_{\mathbf{X}\mathbf{Z}_\mathbf{G}}$ implies that the term $\mathbf{I}_u - \mathbf{S}^{(d)}_{\mathbf{Z}_\mathbf{G}\circ\mathbf{X}}$ equals the sample covariance of the residuals from reduced-rank regression fit of $\mathbf{Z}_\mathbf{G}$ on $\mathbf{X}$ with rank d. Let $\mathrm{F}_{n,1}(\mathbf{G}|u) = \log|\mathbf{G}^T\mathbf{S}_\mathbf{Y}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_\mathbf{Y}^{-1}\mathbf{G}|$ and $\mathrm{F}_{n,2}(\mathbf{G}|d,u) = \log|\mathbf{I}_u - \mathbf{S}^{(d)}_{\mathbf{Z}_\mathbf{G}\circ\mathbf{X}}|$ so that $\mathrm{F}_n(\mathbf{G}|d,u) = \mathrm{F}_{n,1}(\mathbf{G}|u) + \mathrm{F}_{n,2}(\mathbf{G}|d,u)$. The first part $\mathrm{F}_{n,1}(\mathbf{G}|u) \geq 0$ for all $\mathbf{G} \in \mathcal{G}_{r,u}$ and equals zero when $\mathbf{G}$ is a u-dimensional reducing subspace of $\mathbf{S}_\mathbf{Y}$. The effect of $\mathrm{F}_{n,1}(\mathbf{G}|u)$ is then to pull the solution towards eigenvectors of $\mathbf{S}_\mathbf{Y}$. The second part $\mathrm{F}_{n,2}(\mathbf{G}|d,u)$ represents the magnitude of the sample covariance of

the residual from reduced-rank regression fit of the standardized variable $\mathbf{Z}_{\mathbf{G}}$ on $\mathbf{X}$ with given rank d. Simply put, this part is a scale-invariant measure for the lack-of-fit of the rank-d reduced-rank regression of $\mathbf{G}^T\mathbf{Y}$ on $\mathbf{X}$.

Our formulation and decomposition based on (4.3.2) offer a generic way of interpreting the likelihood-based objective functions for envelope methods. For example, the objective function for the standard envelope model in Cook et al. (2010) can be expressed as

$$\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}| + \log|\mathbf{I}_u - \mathbf{S}_{\mathbf{Z}_{\mathbf{G}}\circ\mathbf{X}}|, \tag{4.3.4}$$

which can be interpreted similar to (4.3.2) except the lack-of-fit term is now based on ordinary least squares fit rather than reduced-rank regression fit. The above objective function is the same as (4.3.2) when $d = p$ or $d = u$.

Additional properties of the objective function are given in the following Proposition.

Proposition 4.2. *The objective function* $\mathrm{F}_n(\mathbf{G}|d,u)$ *in (4.3.3) converges in probability as* $n \to \infty$ *to the population objective function* $\mathrm{F}(\mathbf{G}|u) = \log|\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G}| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G}|$ *uniformly in* $\mathbf{G}$. *The estimator* $\widehat{\boldsymbol{\Gamma}} = \arg\min_{\mathbf{G}} \mathrm{F}_n(\mathbf{G}|d,u)$ *is Fisher consistent,* $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathrm{span}\{\arg\min_{\mathbf{G}} \mathrm{F}(\mathbf{G}|u)\}$.

The population objective function $\mathrm{F}(\mathbf{G}|u)$, which does not depend explicitly on the given rank d, is exactly the same one as in Cook et al. (2010) for estimating an u-dimensional envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$. In the proof of Proposition 4.2, we show that $\log[\widehat{\omega}_i(\mathbf{G})]$, for any $i > d$, converges in probability to zero uniformly in $\mathbf{G}$. Therefore, we could view $\mathrm{F}_n(\mathbf{G}|d,u)$ in (4.3.3) as a sample version of $\mathrm{F}(\mathbf{G}|u)$, $\mathrm{F}_n(\mathbf{G}|u) := \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$, plus a finite sample adjustment for the rank deficiency, $\sum_{i=d+1}^{u}\log[\widehat{\omega}_i(\mathbf{G})]$, which goes to zero as $n \to \infty$. Minimizing $\mathrm{F}_n(\mathbf{G}|u)$ leads to another $\sqrt{n}$-consistent envelope estimator but it will not be optimal since it does not account for the rank deficiency. The impact of the rank $d < p$ on the envelope estimation diminishes as sample size increases and reduced-rank envelope estimation moves towards a two-stage estimation procedure: first estimate the envelope from $\mathrm{F}_n(\mathbf{G}|u)$ ignoring the rank, then obtain a rank-d estimator within the estimated envelope. The effects of rank deficiency and envelope interdigitate at finite samples and there is a noticeable synergy when sample size is not large.

Finally, we summarize estimators for the parameters in the reduced-rank envelope model as follows. The results come naturally from Lemma 4.2.

Proposition 4.3. *The estimators for the reduced-rank envelope model (4.2.3) that minimize (4.2.2) are* $\widehat{\boldsymbol{\alpha}}_{\mathrm{RE}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\mathrm{RE}}\overline{\mathbf{X}}$, $\widehat{\boldsymbol{\Gamma}} = \arg\max_{\mathbf{G}\in\mathcal{G}_{r,u}} \mathrm{F}_n(\mathbf{G}|d,u)$, $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\mathbf{S}_{\mathbf{Y}}\widehat{\boldsymbol{\Gamma}}_0$ *and*

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}} &= \mathbf{S}^{1/2}_{\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y}}\left\{\mathbf{I}_u - \mathbf{C}^{(d)}_{\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y},\mathbf{X}}\mathbf{C}^{(d)}_{\mathbf{X},\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y}}\right\}\mathbf{S}^{1/2}_{\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y}} \\
\widehat{\boldsymbol{\Sigma}}_{\mathrm{RE}} &= \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Gamma}}^T + \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T \\
\widehat{\boldsymbol{\beta}}_{\mathrm{RE}} &= \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}\widehat{\mathbf{B}}_{\mathrm{RE}} = \widehat{\boldsymbol{\Gamma}}\mathbf{S}^{1/2}_{\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y}}\mathbf{C}^{(d)}_{\widehat{\boldsymbol{\Gamma}}^T\mathbf{Y},\mathbf{X}}\mathbf{S}^{-1/2}_{\mathbf{X}}.
\end{aligned}$$

The rank of $\widehat{\boldsymbol{\beta}}_{\mathrm{RE}}$ is d and the span of $\widehat{\boldsymbol{\beta}}_{\mathrm{RE}}$ is a subset of the entire u-dimensional envelope. In contrast to reduced-rank regression, the estimator for $\widehat{\boldsymbol{\Sigma}}_{\mathrm{RE}}$ now has an envelope structure:

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{RE}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{RE}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{RE}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}.$$

If we let $u = r$, which is equivalent to setting $\boldsymbol{\Gamma} = \mathbf{I}_r$ in Proposition 4.3, then there is no envelope reduction and the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{RE}}$ is the same as the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{RR}}$ in Lemma 4.1. If we let $d = p$, then the estimators in Proposition 4.3 is the same as the envelope estimators in Cook et al. (2010). The estimators for the reduced-rank envelope model parameters coincide with those estimators in Lemma 4.2 by replacing $\boldsymbol{\Gamma}$ by its estimator $\widehat{\boldsymbol{\Gamma}}$.

4.4 Asymptotics

4.4.1 Asymptotic properties under normality

In this section, we present asymptotic results assuming that the error term is normal, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$, so that the estimators derived in Section 4.3 are all maximum likelihood estimators. We focus attention on the comparison between $\widehat{\boldsymbol{\beta}}_{\mathrm{RE}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{RR}}$ because (1) comparisons between $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ can be found in Cook et al. (2010); and (2) the advantage of $\widehat{\boldsymbol{\beta}}_{\mathrm{RE}}$ over $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ is similar to the advantage of $\widehat{\boldsymbol{\beta}}_{\mathrm{RR}}$ over $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which is due to the rank reduction in the material response $\boldsymbol{\Gamma}^T\mathbf{Y}$. We then relax the normality assumption in Section 4.4.2 and show the $\sqrt{n}$-consistency of the reduced-rank envelope estimator and its asymptotic distribution.

From Cook et al. (2010) we know that the Fisher information for $\mathbf{h}$ is

$$\mathbf{J}_{\mathbf{h}} = \begin{pmatrix} \mathbf{J}_{\boldsymbol{\beta}} & 0 \\ 0 & \mathbf{J}_{\boldsymbol{\Sigma}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix}, \tag{4.4.1}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \mathbf{S}_{\mathbf{X}}$ and $\mathbf{E}_r$ is the expansion matrix, $\mathbf{E}_r \mathrm{vec}(\mathbf{S}) = \mathrm{vech}(\mathbf{S})$ for any $r \times r$ symmetric matrix $\mathbf{S}$. The asymptotic covariance for the ordinary least squares estimator $\widehat{\mathbf{h}}_{\mathrm{OLS}}$ is $\mathbf{J}_{\mathbf{h}}^{-1}$, which is the asymptotic covariance of the unrestricted maximum likelihood estimator.

Define the gradient matrices

$$\mathbf{H} = \frac{\partial \mathbf{h}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \text{ and } \mathbf{R} = \frac{\partial \mathbf{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}. \tag{4.4.2}$$

Then the asymptotic covariance for the reduced-rank regression estimator $\widehat{\mathbf{h}}_{\mathrm{RR}} = \mathbf{h}(\widehat{\boldsymbol{\psi}})$ and for the reduced-rank envelope estimator $\widehat{\mathbf{h}}_{\mathrm{RE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ are summarized in the following Proposition.

Proposition 4.4. *Assuming that* $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$, *then* $\text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{OLS}}) = \mathbf{J}_{\mathbf{h}}^{-1}$, $\text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{RR}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T$ *and* $\text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{RE}}) = \mathbf{R}(\mathbf{R}^T\mathbf{J}_{\mathbf{h}}\mathbf{R})^{\dagger}\mathbf{R}^T$. *Moreover,*

$$\begin{aligned}
\text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{OLS}}) - \text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{RR}}) &= \mathbf{J}_{\mathbf{h}}^{-1/2}\mathbf{Q}_{\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{H}}\mathbf{J}_{\mathbf{h}}^{-1/2} \geq 0, \\
\text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{RR}}) - \text{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{RE}}) &= \mathbf{J}_{\mathbf{h}}^{-1/2}\left(\mathbf{P}_{\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{R}}\right)\mathbf{J}_{\mathbf{h}}^{-1/2} \\
&= \mathbf{J}_{\mathbf{h}}^{-1/2}\mathbf{P}_{\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{H}}\mathbf{Q}_{\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{R}}\mathbf{J}_{\mathbf{h}}^{-1/2} \geq 0,
\end{aligned}$$

where † *indicates the Moore-Penrose inverse. In particular,*

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})] \geq \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] \geq \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})].$$

Proposition 4.4 follows directly from $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\phi})$. Therefore, we have $\mathbf{R} = \mathbf{H}\partial\boldsymbol{\psi}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$ and $\text{span}(\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{R}) \subseteq \text{span}(\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{H})$. Similarly, it can be shown that $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})] \geq \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{env}})] \geq \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$.

Since we are particularly interested in the asymptotic covariance of $\widehat{\mathbf{h}}_1 = \text{vec}(\widehat{\boldsymbol{\beta}})$ from different estimators, we summarize some of the results in the following Propositions.

Proposition 4.5. *Assume that* $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ *and that* $\text{rank}(\boldsymbol{\beta}) = d$. *Then* $\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})$ *and* $\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}} - \boldsymbol{\beta})$ *are both asymptotically normal with mean zero and covariances as follows.*

$$\begin{aligned}
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})] &= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} \\
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] &= (\mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})})\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})] \quad (4.4.3) \\
&= \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}}\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}^T)] + \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{B}})] \quad (4.4.4)
\end{aligned}$$

where $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}})] = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\boldsymbol{\Sigma})$ *and* $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{B}})] = (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}) \otimes \boldsymbol{\Sigma}$.

The asymptotic result in (4.4.3) follows from Anderson (1999; equation (3.20)). The results in Proposition 4.5 rely on $\mathbf{A}$ and $\mathbf{B}$ only through their projections $\mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}$ and $\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}$, which serve to orthogonalize the parameters in the asymptotic variance decompositions. This implies that all the equalities in Proposition 4.5 hold for any decomposition $\boldsymbol{\beta} = \mathbf{AB}$, with $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times p}$. Hence, Proposition 4.5 is a unification for all the asymptotic studies of reduced-rank regression in the literature such as Anderson (1999), Reinsel and Velu (1998), Stoica and Viberg (1996) and so on.

For the reduced-rank envelope model (4.2.3), we have the following results on asymptotic distributions.

Proposition 4.6. *Under the reduced-rank envelope model with normal error* $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$, $\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}} - \boldsymbol{\beta})$ *is asymptotically normal with mean zero and covariance*

$$\begin{aligned}
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})] &= \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] + \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})] \\
&= \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}}\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}^T)] + \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\mathbf{B}})] \\
&\quad + \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})], \quad (4.4.5)
\end{aligned}$$

where $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}}\mathbf{Q}^T_{\mathbf{B}^T(\boldsymbol{\Sigma}_\mathbf{X})})] = \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_\mathbf{A}\mathbf{Q}^T_{\mathbf{B}^T(\boldsymbol{\Sigma}_\mathbf{X})})]$ *from (4.4.4). Explicit expressions for* $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$ *can be found in the Supplemental material (4.9.27). The above equalities hold for any decomposition* $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{B}$*, where* $\boldsymbol{\Gamma}$ *is semi-orthogonal and the dimensions of* $\boldsymbol{\Gamma}$*,* $\boldsymbol{\eta}$ *and* $\mathbf{B}$ *are* $r \times u$*,* $u \times d$ *and* $d \times p$*.*

We view the asymptotic advantages of reduced-rank envelopes over reduced-rank regression by contrasting (4.4.4) with (4.4.5). From Propositions 4.5 and 4.6, we can write $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] - \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$ as

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_\mathbf{B})] - \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\mathbf{B}})] - \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})] \geq 0, \tag{4.4.6}$$

where $\widehat{\boldsymbol{\beta}}_\mathbf{B} = \widehat{\mathbf{A}}_\mathbf{B}\mathbf{B}$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\mathbf{B}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma},\mathbf{B}}\mathbf{B} = \widehat{\mathbf{A}}_{\boldsymbol{\Gamma},\mathbf{B}}\mathbf{B}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}} = \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta},\mathbf{B}}\boldsymbol{\eta}\mathbf{B} = \widehat{\mathbf{A}}_{\boldsymbol{\eta},\mathbf{B}}\mathbf{B}$ are estimators with given $\mathbf{B}$. When $\mathbf{B}$ is known, the original regression problem simplifies to the regression of $\mathbf{Y}$ on $\mathbf{BX}$ and $\mathbf{A}$ is the new regression coefficient matrix. The estimator $\widehat{\mathbf{A}}_\mathbf{B}$ is the ordinary least squares estimator of $\mathbf{Y}$ on $\mathbf{BX}$ and the estimators $\widehat{\mathbf{A}}_{\boldsymbol{\Gamma},\mathbf{B}}$ and $\widehat{\mathbf{A}}_{\boldsymbol{\eta},\mathbf{B}}$ correspond to the usual envelope estimators for $\mathbf{A} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. The difference in asymptotic covariances $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] - \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$ from (4.4.6) equals the asymptotic efficiency gain of envelope estimator over the ordinary least squares estimator for regression of $\mathbf{Y}$ on $\mathbf{BX}$ and is consistent with the results presented in Cook et al. (2010).

Two special situations where the inequality in (4.4.6) becomes equality are: $\boldsymbol{\Gamma} = \mathbf{I}_r$ and $\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2\mathbf{I}_r$, while the envelope estimator is asymptotically equivalent to the ordinary least squares estimator in these two cases.

To see the potential gain of the reduced-rank envelope estimator, we have the following Corollary, where we have ignored the cost of estimating an envelope.

Corollary 4.1. *Under the reduced-rank envelope model with normal error* $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$*,*

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] = \mathbf{F}_1\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{env},\boldsymbol{\Gamma}})] = \mathbf{F}_2\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] = \mathbf{F}_1\mathbf{F}_2\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})],$$

where $\mathbf{F}_1 = \mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_\mathbf{X})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}$ *and* $\mathbf{F}_2 = \mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}}$ *are two positive semi-definite matrices with eigenvalues between 0 and 1.*

The two matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ represent fractions of asymptotic covariance reduction from the ordinary least squares estimator to the reduced-rank regression estimator and to the envelope estimator with given $\boldsymbol{\Gamma}$. Then the efficiency gain of reduced-rank envelope with known $\boldsymbol{\Gamma}$ over ordinary least squares is the superimposition of the efficiency gain of the reduced-rank regression and the envelope regression with known $\boldsymbol{\Gamma}$.

4.4.2 Consistency without the normality assumption

Let $\widehat{\mathbf{h}}_{\text{OLS}} = \left(\text{vec}^T(\widehat{\boldsymbol{\beta}}_{\text{OLS}}), \text{vech}^T(\mathbf{S}_{\mathbf{Y}|\mathbf{X}})\right)^T$ denote the ordinary least squares estimator of $\mathbf{h}$ under the standard linear regression model, and let $\widehat{\mathbf{h}}_{\text{RE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ denote the reduced-rank

envelope estimator. The true values of $\mathbf{h}$ and $\boldsymbol{\phi}$ are denoted as $\mathbf{h}_0$ and $\boldsymbol{\phi}_0$. The objective function $L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ in (4.2.2) can be written as, after partially maximized over $\boldsymbol{\alpha}$,

$$L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \simeq -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\text{trace}\left\{\boldsymbol{\Sigma}^{-1}[\mathbf{S}_{\mathbf{Y}|\mathbf{X}} + (\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})\mathbf{S}_{\mathbf{X}}(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})^T]\right\}. \quad (4.4.7)$$

We treat the objective function $L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ as a function of $\mathbf{h}$ and $\widehat{\mathbf{h}}_{\text{OLS}}$ and define $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}}) = 2/n\left\{L_n(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \mathbf{S}_{\mathbf{Y}|\mathbf{X}}) - L_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\}$, which satisfies the conditions of Shapiro's (1986) minimum discrepancy function (see Supplement Section 4.9.6). Hence $\mathbf{J}_{\mathbf{h}}$ is equal to $1/2 \times \partial^2\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}})/\partial\mathbf{h}\partial\mathbf{h}^T$ evaluated at $\widehat{\mathbf{h}}_{\text{OLS}} = \mathbf{h} = \mathbf{h}_0$ is the Fisher information matrix for $\mathbf{h}$ when $\boldsymbol{\epsilon}$ is normal. The following proposition formally states the asymptotic distribution of $\widehat{\mathbf{h}}_{\text{RE}}$ without normality of $\boldsymbol{\epsilon}$.

Proposition 4.7. *Assume that the reduced-rank envelope model (4.2.3) holds and that $\boldsymbol{\epsilon}_i$'s are independent and identically distributed with finite fourth moments. Then $\sqrt{n}(\widehat{\mathbf{h}}_{\text{OLS}} - \mathbf{h}_0) \to N(0, \mathbf{K})$, for some positive definite covariance matrix $\mathbf{K}$. And $\sqrt{n}(\widehat{\mathbf{h}}_{\text{RE}} - \mathbf{h}_0)$ converges in distribution to a normal random variable with mean $\mathbf{0}$ and covariance matrix*

$$\mathbf{W} = \mathbf{R}\left(\mathbf{R}^T\mathbf{J}_{\mathbf{h}}\mathbf{R}\right)^{\dagger}\mathbf{R}^T\mathbf{J}_{\mathbf{h}}\mathbf{K}\mathbf{J}_{\mathbf{h}}\mathbf{R}\left(\mathbf{R}^T\mathbf{J}_{\mathbf{h}}\mathbf{R}\right)^{\dagger}\mathbf{R}^T.$$

In particular, $\sqrt{n}(\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}}) - \text{vec}(\boldsymbol{\beta}))$ converges in distribution to a normal random variable with mean $\mathbf{0}$ and covariance $\mathbf{W}_{11}$, the upper-left $pr \times pr$ block of $\mathbf{W}$. The explicit expression for the gradient matrix $\mathbf{R} = \partial\mathbf{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$ is given in the Supplement equation (4.9.21).

The $\sqrt{n}$-consistency of the reduced-rank envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{RE}}$ is essentially because that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ are $\sqrt{n}$-consistent regardless of normality assumption and also because of the properties of $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}})$. The asymptotic covariance matrix $\mathbf{W}_{11}$ can be estimated straightforwardly using the plug-in method once $\mathbf{K}$ is estimated, but its accuracy for any fixed sample size will depend on the distribution of $\boldsymbol{\epsilon}$, which is usually unknown in practice. Fortunately, bootstrap methods can provide good estimates of $\mathbf{W}_{11}$, as illustrated in Section 4.6.3.

4.5 Selections of rank and envelope dimension

4.5.1 Rank

We apply the same testing procedure as in Section 3.5.1, which was proposed by Bura and Cook (2003). They developed a chi-squared test for the rank d that requires only that the response variables have finite second moments. The test statistic is $\Lambda_d = n\sum_{j=d+1}^{\min(p,r)}\varphi_j^2$, where $\varphi_1 \geq \cdots \geq \varphi_{\min(p,r)}$ are eigenvalues of the $p \times r$ matrix

$$\widehat{\boldsymbol{\beta}}_{\text{std}} = \{(n-p-1)\}/n\}^{1/2}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}\widehat{\boldsymbol{\beta}}_{\text{OLS}}\mathbf{S}_{\mathbf{X}}^{1/2}. \quad (4.5.1)$$

Under the null hypothesis that $H_0 : d = d_0$, Bura and Cook (2003) showed that Λ_{d_0} is asymptotically distributed as a $\chi^2_{(p-d_0)(r-d_0)}$ random variable. The rank d is then determined by comparing a sequence of test statistics Λ_{d_0}, $d_0 = 0, \ldots, \min(p, r) - 1$, to the percentiles of their null distribution $\chi^2_{(p-d_0)(r-d_0)}$. The sequence of tests terminates at the first non-significant test of $H_0 : d = d_0$ and then d_0 serves as an estimate of the rank of $\boldsymbol{\beta}$.

4.5.2 Envelope dimension

Since the envelope dimension satisfies $d \leq u \leq r$, standard techniques such as sequential likelihood-ratio tests, AIC and BIC can be applied to select u, as in Cook et al. (2010).

For any possible combination (d, u) with $0 \leq d \leq u \leq r$, let $\widehat{L}_{d,u}$ denote the maximized log-likelihood function (c.f. (4.9.3)), which is evaluated at the maximum likelihood estimators in Proposition 4.3. Assuming d is known, then $\Lambda_{d,u_0} = 2(\widehat{L}_{d,r} - \widehat{L}_{d,u_0})$ is asymptotically distributed as a $\chi^2_{(r-u_0)d}$ random variable under the null hypothesis $H_0 : u = u_0$. Thus, a sequence of likelihood ratio tests of $u_0 = d, \ldots, r-1$ can be used to determine u after d is determined by the method described in Section 4.5.1. The first non-significant value of u_0 will serve as the envelope dimension.

Information criteria such as AIC and BIC can be used to select (d, u) simultaneously. We write AIC as $\mathcal{A}_{d,u} = 2K_{d,u} - 2\widehat{L}_{d,u}$, where $K_{d,u} = (p + u - d)d + r(r+1)/2$ is the total number of parameters in the reduced-rank envelope model, and write BIC as $\mathcal{B}_{d,u} = \log(n)K_{d,u} - 2\widehat{L}_{d,u}$. We search (d, u) from $(0, 0)$ to (r, r) with constraint $d \leq u$ and choose the pair that has the smallest AIC or BIC. Alternatively, we can first determine d from the asymptotic chi-squared tests in Section 4.5.1 and then search u from $d, \ldots, r$ with the smallest AIC or BIC, which could save a lot of computation. The computation cost for determining d by the sequential chi-squared tests in Section 4.5.1 is substantially cheaper than the computation cost in calculating AIC and BIC, which involves sequence of Grassmannian optimizations.

When sample size is not too small, our experience suggests that the most favorable procedure is BIC selection for $u = d, \ldots, r$ where d is guided by the sequential chi-squared tests. Since the true envelope dimension always exist, BIC is consistent in the sense that the probability of selecting the correct u approaches 1, given the correct d. There are many articles comparing AIC and BIC, from both theoretical and practical points of view, for example Shao (1997) and Yang (2005).

The rank d and envelope dimension u can also be determined by cross-validation or by using hold-out samples. These approaches are especially appropriate when prediction is the primary goal of the study rather than correctness of the selected model.

4.6 Simulations

4.6.1 Rank and dimension

In all the simulations, we first filled in $\mathbf{\Gamma}$, $\boldsymbol{\eta}$ and $\mathbf{B}$ with random uniform (0,1) numbers, and then $\mathbf{\Gamma}$ was standardized so that $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$ and $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{B}$ was standardized so that $||\boldsymbol{\beta}||_F = 1$. Estimation errors were defined as $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$. Unless otherwise specified, the predictors and errors were simulated independently from $N(0, \mathbf{I}_p)$ and $N(0, \mathbf{\Sigma})$ distributions. All figures were generated based on averaging over 200 independent replicate data sets.

In this section, we present simulation results to demonstrate the behavior of the proposed method using various sample sizes, ranks d, dimensions u. We simulated data from model (4.2.3), where $[\mathbf{\Omega}]_{ij} = (-0.9)^{|i-j|}$ and $[\mathbf{\Omega}_0]_{ij} = 5 \cdot (-0.5)^{|i-j|}$. Figure 4.1 summarizes the effect of dimension and rank on the relative performances of each methods. In the left plot $(d, u, p, r) = (1, 10, 10, 20)$. Since the rank was only one but the envelope dimension was ten, reduced-rank regression had a dramatic improvement over ordinary least squares, while the ordinary envelope method had a relatively modest gain over ordinary least squares. The reduced-rank envelope had a relatively small edge over reduced-rank regression. The second case was $(d, u, p, r) = (4, 5, 6, 20)$, where $\boldsymbol{\beta}$ had nearly full column rank and the envelope dimension was much smaller than the number of response variables. Not surprisingly, reduced-rank regression had modest gain over ordinary least squares while the envelope estimator and the reduced-rank envelope estimator had similar behavior and significantly improved over ordinary least squares and reduced-rank regression. The last case was chosen as $(d, u, p, r) = (5, 10, 15, 20)$ so that there was no particular favor towards either the envelope method or reduced-rank regression. We found good improvement over ordinary least squares by both reduced-rank regression and envelopes. However, reduced-rank envelopes combined both of their strengths and resulted in a bigger gain.

We found in practice that reduced-rank envelopes typically have improved performance over reduced-rank regression and envelope estimators, and it has similar behavior to one of the two estimators if the other one performed poorly. Even in the extreme cases where $d = p$ or $u = r$, reduced-rank envelopes can still gain drastically over ordinary least squares similar to the results in Figure 4.1.

We next illustrate the asymptotic chi-squared test for rank detection combined with BIC selection for envelope dimension, as discussed in Section 4.5. Using the same simulation model, we took $(d, u, p, r) = (3, 5, 8, 12)$, where the total number of parameters in the reduced-rank envelope model was 108. The percentages of correct detections for d an u were plotted in Figure 4.2 versus sample size. The BIC selection of u was based on the correct rank d. The significance level of the chi-squared tests was 0.05. As seen from the figure, the probability of selecting the correct d was about 0.9 at $n = 400$ samples and the probability of correct detection settled at 95% for larger n, as predicted by the

hypothesis testing theory. BIC selection for the envelope dimension u seemed to be very accurate even with small samples. The likelihood-ratio tests and AIC selection for u were not nearly as effective as BIC and thus were omitted from the plot. We also considered BIC selection for u and d simultaneously. The probability of simultaneous correctness was less than 70% for $n \leq 600$ but reached more than 95% correctness for $n \geq 900$. In our experience the best method for determining dimensions is to use the chi-squared test for d and BIC selection on u based on the selected d. Overestimation of d and u usually is not a serious issue but underestimation of d and u will certainly cause bias in estimation.

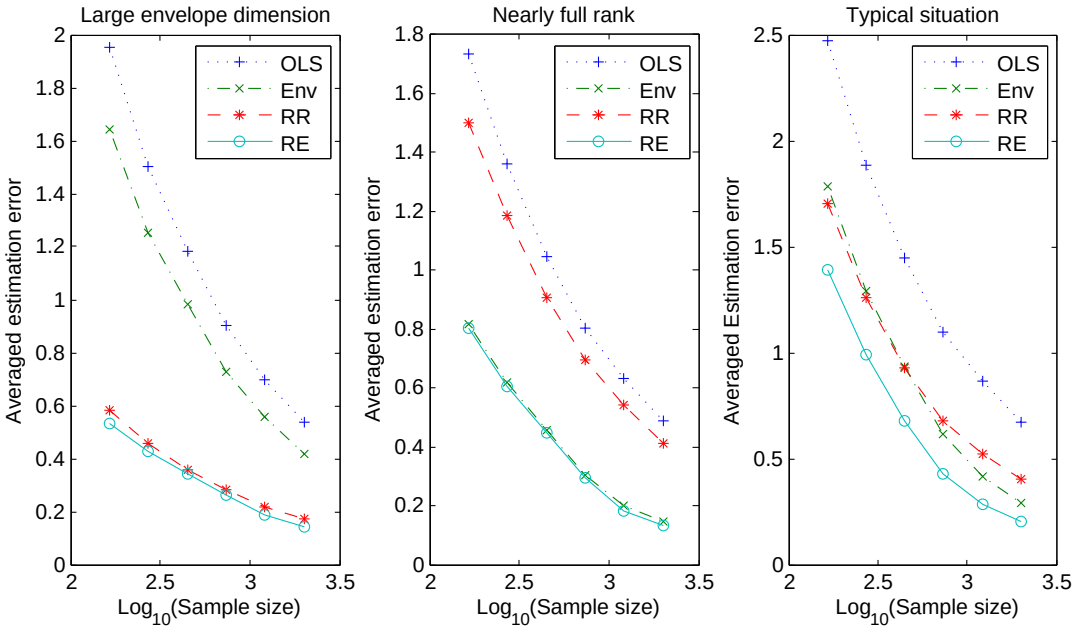


Figure 4.1: Effect of rank and dimension. Averaged estimation error on the vertical axis is defined as averaged $||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}||_F$ over 200 independent data sets. The dimensions of the three plots were: (1) large envelope dimension case $(d, u, p, r) = (1, 10, 10, 20)$; (2) nearly full rank case $(d, u, p, r) = (4, 5, 6, 20)$ and (3) a typical situation $(d, u, p, r) = (5, 10, 15, 20)$. The sample sizes varied from 160 to 2000 and were shown in a logarithmic scale.

4.6.2 Signal-versus-noise and material-versus-immaterial

In this section, we describe the behavior of each method with varying signal-to-noise ratios and ratios of immaterial variation to material variation. We fixed the sample size at 400 and the dimensions were $(d, u, p, r) = (3, 7, 10, 20)$. The covariances had the forms of $[\boldsymbol{\Omega}]_{ij} = \sigma^2 \cdot (-0.9)^{|i-j|}$ and $[\boldsymbol{\Omega}_0]_{ij} = \sigma_0^2 \cdot (-0.9)^{|i-j|}$ with varying constants $\sigma^2,\ \sigma_0^2 > 0$.

In the study of varying signal-to-noise ratio, we kept $\sigma^2 = \sigma_0^2$. And because $||\boldsymbol{\beta}||_F = 1$, the signal-to-noise ratio was simply $1/\sigma^2$ which varied from 0.1 to 10. Figure 4.3 summarizes the results of two numerical experiments. All the four lines in this log-log scale signal-to-noise ratio plot are roughly parallel, which implies that the four methods are

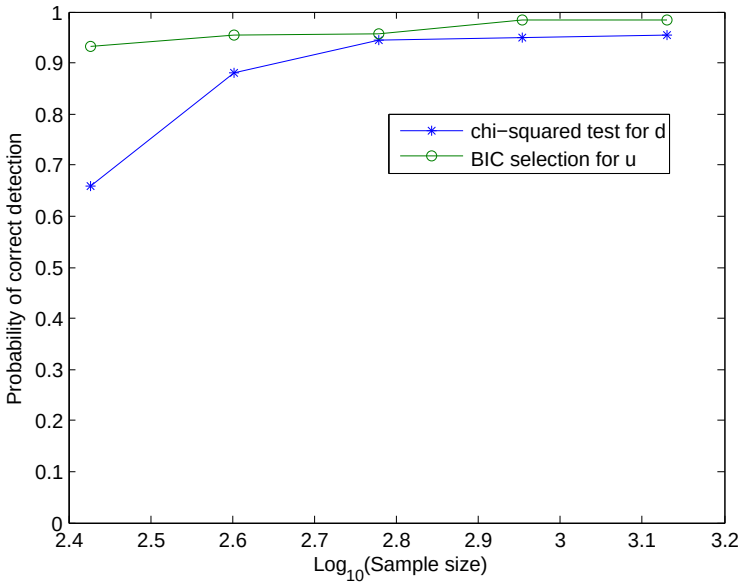


Figure 4.2: The empirical probability of correct detection versus sample sizes. BIC selection on u was based on true rank d.

exponentially more distinguishable in weaker signal. Comparing reduced-rank regression to envelopes, the reduced-rank regression seemed to perform better in stronger signals (signal-to-noise ratio$\geq$ 1), but the envelope estimator was less vulnerable to weaker signals (signal-to-noise ratio$\leq$ 1). This was because the envelope method can gain information from the error term $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ while reduced-rank regression and the standard method cannot. Reduced-rank envelope estimators combined the strengths of reduced-rank regression and envelopes, and hence outperformed both estimators in strong and weak signals.

In the study of varying immaterial-to-material variance ratio, we kept $\sigma^2 = 1$ and changed σ_0^2. The ratio is then defined as σ_0^2 and the horizontal axis in the plot is $\log_{10}(\sigma_0^2)$, which varied from -0.5 to 2. Not surprisingly, reduced-rank regression and ordinary least squares behaved similarly because they did not gain information from the covariance structure of $\boldsymbol{\Sigma}$. The envelope estimator and the reduced-rank envelope estimator had similar behavior, and they had much better performances over ordinary least squares and reduced-rank regression when the immaterial variation was large. This is due to the fact that envelope methods can efficiently eliminate the immaterial information. In this example, the averaged estimation errors for ordinary least squares, reduced-rank regression and envelope were 7.2, 3.9 and 1.8 times of that of the envelope reduced-rank regression when $\sigma_0^2 = 100$.

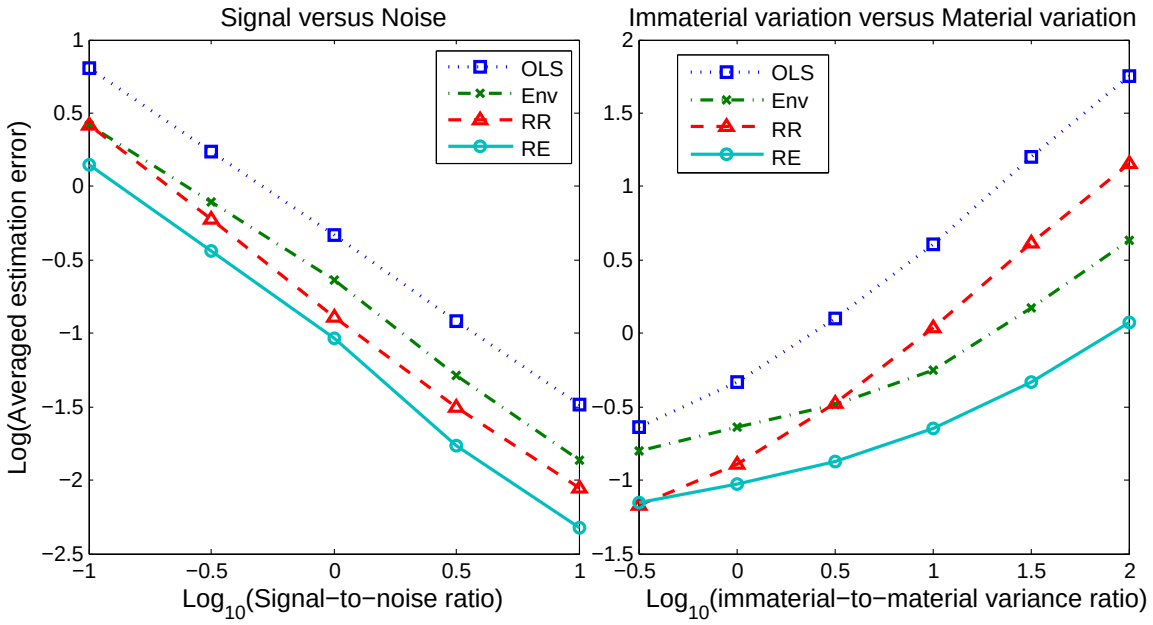


Figure 4.3: Varying the signal-to-noise ratio and the immaterial-to-material variance ratio.

4.6.3 Bootstrap standard errors

To illustrate the application of the bootstrap for estimating the standard errors of regression coefficients, we considered a model with $(d, u, p, r) = (2, 4, 6, 8)$. Residual bootstrap samples were used since we considered $\mathbf{X}$ as a non-stochastic predictor. Both $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ were randomly generated as $\mathbf{M}\mathbf{M}^T$, where $\mathbf{M} \in \mathbb{R}^{4\times 4}$ was filled with uniform (0,1) numbers. The error term $\boldsymbol{\epsilon}_i$ was simulated as $\boldsymbol{\epsilon}_i = \boldsymbol{\Sigma}^{1/2}\mathbf{U}_i$, where $\mathbf{U}_i$ was a vector of i.i.d. random variable with mean 0 standard deviation 1. We simulated both normal and uniform $\mathbf{U}_i$. The standard errors of a selected element in $\widehat{\boldsymbol{\beta}}$ were plotted in Figure 4.4. For both normal and non-normal data, the three types of standard error estimates agreed well: the theoretical standard errors were the squared roots of the diagonal elements in the asymptotic covariances of each estimators divided by $\sqrt{n}$; the actual standard errors were based on 200 independent realizations; and the bootstrap standard errors were based on 200 bootstrap replicate data sets. Moreover, the bootstrap standard errors were close to the theoretical standard errors of the maximum likelihood estimators even when the normality assumption was violated. As expected, the reduced-rank envelope estimator had much smaller standard errors than those of the ordinary least squares estimator. We also simulated non-normal errors from t-distribution and χ^2-distribution, and obtained results similar to Figure 4.4.

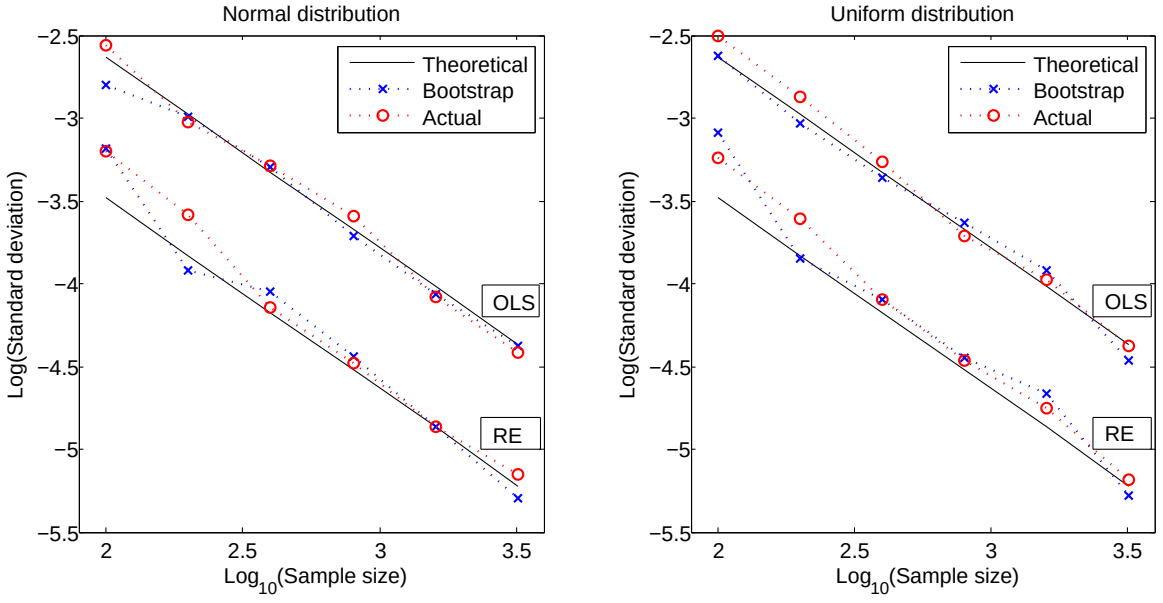


Figure 4.4: Theoretical, bootstrap and actual standard errors with normal and uniform errors $\boldsymbol{\epsilon}$. The sample sizes were 100, 200, 400,. . . , 3200. The standard errors for reduced-rank regression and envelope estimators were consistently between the ordinary least squares and the envelope reduced-rank regression standard errors, and were not included in these plots for better visualization.

4.7 Sales people test scores data

This data set consisted of 50 sales people from a firm. Three performances variables were used as predictors: growth of sales (X_1), profitability of sales (X_2) and new account sales (X_3). And four response variables were test scores on creativity (Y_1), mechanical reasoning (Y_2), abstract reasoning (Y_3) and mathematical ability (Y_4). This data set can be found in Johnson and Wichern (2007).

The chi-squared rank test in Section 4.5.1 suggested that $d = 2$ at level 0.01. Then based on BIC we selected the envelope dimension to be $u = 3$. We computed the fractions $f_{ij} := 1 - \widehat{\text{avar}}^{1/2}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{ij,\text{RE}})/\widehat{\text{avar}}^{1/2}(\sqrt{n}\widetilde{\boldsymbol{\beta}}_{ij})$ for all i and j, where $\widetilde{\boldsymbol{\beta}}$ denotes one of the estimators to be compared: $\widehat{\boldsymbol{\beta}}_{\text{RR}}$, $\widehat{\boldsymbol{\beta}}_{\text{env}}$ and $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$. Comparing to ordinary least squares , the standard deviations of the elements in the reduced-rank envelope estimator were 5% to 60% smaller, $0.05 \leq f_{ij} \leq 0.60$. Hypothetically, a sample size of more than 300 observations, in contrast to the original 50 observations, would be needed to achieve a 60% smaller standard deviation in ordinary least squares. The fractions for comparing with the reduced-rank regression estimator were $0.01 \leq f_{ij} \leq 0.24$, where 24% smaller standard

deviation than reduced-rank regression implies a doubling of the observations for reduced-rank regression to achieve the same performance as reduced-rank envelope estimator. At last, the reduced-rank envelope estimator compared to the ordinary envelope estimator, had 3% to 51% smaller standard deviations, where 51% smaller standard deviation meant four times the sample size, $n = 200$, for the ordinary envelope estimator.

4.8 Discussion

The predictor envelope model (Cook et al. 2013) is a sibling of the envelope model (Cook et al. 2010). In that model, joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is assumed, which led to a likelihood analysis for the predictor envelope $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$, where $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the regression coefficient matrix in:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{4.8.1}$$

and $\boldsymbol{\epsilon}$ is normally distributed and independent of $\mathbf{X}$. Therefore, we can parameterize $\boldsymbol{\beta} = \boldsymbol{\xi}\mathbf{\Gamma}^T$ and $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Phi}_0\mathbf{\Gamma}_0^T$, where $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$ is some semi-orthogonal basis matrix of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}^T)$. Since the predictor envelope focuses on reducing $\mathbf{X}$, it can gain efficiency in estimating $\boldsymbol{\beta}$ and in prediction even when $\mathbf{Y}$ is univariate.

It is easy to extend predictor envelope to the reduced-rank envelope model. Assuming that $\text{rank}(\boldsymbol{\beta}) = d \leq u$, then we could write $\boldsymbol{\beta} = \mathbf{Ab\Gamma}^T$ for some matrices $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{b} \in \mathbb{R}^{d\times u}$. Then by partially maximizing the likelihood, similar to the derivation in Section 4.3.2, we could get the likelihood-based objective function for estimating the predictor envelope

$$\mathrm{F}_n(\mathbf{G}|d,u) = \log|\mathbf{G}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}| + \sum_{i=d+1}^{u}\log|\widehat{\omega}_i(\mathbf{G})|. \tag{4.8.2}$$

where $\widehat{\omega}_i(\mathbf{G})$ is the i-th singular value of $(\mathbf{G}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{S}_{\mathbf{X}}\mathbf{G})$. Then after obtaining $\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{G}\in\mathcal{G}_{p,u}}\mathrm{F}_n(\mathbf{G}|d,u)$, the likelihood-based estimators for $\mathbf{A}$ and $\mathbf{b}$ is the same as the reduced-rank regression estimator of $\mathbf{Y}$ on $\widehat{\mathbf{\Gamma}}^T\mathbf{X}$. Therefore, extension of the reduced-rank envelope model to predictor space could improve estimation accuracy over reduced-rank regression when $r = d < u < p$ that reduced-rank regression has no gain over ordinary least squares .

4.9 Proofs and technical details

4.9.1 Maximizing the likelihood-based objective function (4.2.2)

In this Section, we consider maximizing $L_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Sigma})$ from (4.2.2) under different model parameterizations regarding standard, reduced-rank, envelope and reduced-rank envelope

models. Maximizing L_n from (4.2.2) is equivalent to deriving maximum likelihood estimators with normally distributed error $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ as follows. Lemmas 4.1 and 4.2 and Propositions 4.1 and 4.3 are proved directly in the derivation of estimators.

Standard regression and envelope regression

Maximum likelihood estimators for the standard regression model is the ordinary least squares estimator, $\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{OLS}} = \mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. From Cook et al. (2010), we have the maximum likelihood estimators for the envelope model as

$$\begin{aligned}
\widehat{\boldsymbol{\Gamma}}_{\text{env}} &= \arg\min_{\mathbf{G}\in\mathcal{G}_{r,u}} \left\{\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|\right\} \\
\widehat{\boldsymbol{\beta}}_{\text{env}} &= \widehat{\boldsymbol{\Gamma}}_{\text{env}}\mathbf{S}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}^T\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}}\widehat{\boldsymbol{\beta}}_{\text{OLS}} \\
\widehat{\boldsymbol{\Sigma}}_{\text{env}} &= \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}}\mathbf{S}_{\mathbf{Y}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}_{\text{env}}}.
\end{aligned}$$

Reduced-rank regression (i.e., proof of Lemma 4.1)

Following Anderson (1999) equation (2.13), we let $\widehat{\mathbf{L}} \in \mathbb{R}^{p\times d}$ denote $\mathbf{S}_{\mathbf{X}}^{-1/2}[\mathbf{v}_1, \ldots, \mathbf{v}_d]$, where $\mathbf{v}_i$ is the i-th eigenvector of $\mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{S}_{\mathbf{X}\circ\mathbf{Y}}\mathbf{S}_{\mathbf{X}}^{-1/2}$. Then the estimators can be written as $\widehat{\boldsymbol{\alpha}}_{\text{RR}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\text{RR}}\overline{\mathbf{X}}$, $\widehat{\boldsymbol{\beta}}_{\text{RR}} = \mathbf{S}_{\mathbf{YX}}\widehat{\mathbf{L}}\widehat{\mathbf{L}}^T$ and $\widehat{\boldsymbol{\Sigma}}_{\text{RR}} = \mathbf{S}_{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\text{RR}}\mathbf{S}_{\mathbf{XY}}$. We then use the sample canonical correlation matrix notion to get the results in Lemma 4.1: $\mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{S}_{\mathbf{X}\circ\mathbf{Y}}\mathbf{S}_{\mathbf{X}}^{-1/2} = \mathbf{C}_{\mathbf{XY}}\mathbf{C}_{\mathbf{YX}}$ and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{RR}} &= \mathbf{S}_{\mathbf{YX}}\widehat{\mathbf{L}}\widehat{\mathbf{L}}^T = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{X}}^{-1/2}\mathbf{P}_{\mathbf{C}_{\mathbf{XY}}^{(d)}}\mathbf{S}_{\mathbf{X}}^{-1/2} \\
&= \mathbf{S}_{\mathbf{Y}}^{1/2}\mathbf{C}_{\mathbf{YX}}\mathbf{P}_{\mathbf{C}_{\mathbf{XY}}^{(d)}}\mathbf{S}_{\mathbf{X}}^{-1/2} = \mathbf{S}_{\mathbf{Y}}^{1/2}\mathbf{C}_{\mathbf{YX}}^{(d)}\mathbf{S}_{\mathbf{X}}^{-1/2}.
\end{aligned}$$

Reduced-rank envelope regression

Proof of Lemma 4.2 Estimation for the envelope model is facilitated by the following consideration which is straightforward from (4.2.3).

$$\boldsymbol{\Gamma}^T\mathbf{Y}_i = \boldsymbol{\Gamma}^T\boldsymbol{\alpha} + \boldsymbol{\eta}\mathbf{B}\mathbf{X}_i + \boldsymbol{\Gamma}^T\boldsymbol{\epsilon}_i, \tag{4.9.1}$$

$$\boldsymbol{\Gamma}_0^T\mathbf{Y}_i = \boldsymbol{\Gamma}_0^T\boldsymbol{\alpha} + \boldsymbol{\Gamma}_0^T\boldsymbol{\epsilon}_i, \tag{4.9.2}$$

where $\boldsymbol{\Gamma}^T\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Omega})$, $\boldsymbol{\Gamma}_0^T\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Omega}_0)$, $\boldsymbol{\Gamma}^T\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\boldsymbol{\epsilon}$.

The maximum likelihood estimator of $\boldsymbol{\alpha}$ is $\widehat{\boldsymbol{\alpha}}_{\text{RE}} = \overline{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}_{\text{RE}}\overline{\mathbf{X}}$ and effectively we could use centered response $\mathbf{Y}_{ci} := \mathbf{Y}_i - \overline{\mathbf{Y}}$ and centered predictors $\mathbf{X}_{ci} = \mathbf{X}_i - \overline{\mathbf{X}}$ to omit the analysis on $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}_{\text{RE}}$. Then the partially maximized log-likelihood with known dimensions u and d can be decomposed into the following two additive parts since $\boldsymbol{\Gamma}^T\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\Gamma}_0^T\boldsymbol{\epsilon}$.

$$L_n(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}|d, u) \simeq L_{1,n}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}|d, u) + L_{2,n}(\boldsymbol{\Gamma}_0, \boldsymbol{\Omega}_0|u) \tag{4.9.3}$$

where $L_{1,n}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}|d, u)$ corresponds to the likelihood from (4.9.1) and is given by

$$-\frac{n}{2}\left\{\log|\boldsymbol{\Omega}| + \text{trace}\left[\boldsymbol{\Omega}^{-1}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^T\mathbf{Y}_{ci} - \boldsymbol{\eta}\mathbf{B}\mathbf{X}_{ci})(\boldsymbol{\Gamma}^T\mathbf{Y}_{ci} - \boldsymbol{\eta}\mathbf{B}\mathbf{X}_{ci})^T\right]\right\}, \tag{4.9.4}$$

and $L_{2,n}(\boldsymbol{\Gamma}_0, \boldsymbol{\Omega}_0|u)$ corresponds to the likelihood from (4.9.2) and is equal to

$$-\frac{n}{2}\left\{\log|\boldsymbol{\Omega}_0| + \text{trace}\left[\boldsymbol{\Omega}_0^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}_0^T\mathbf{Y}_{ci}\mathbf{Y}_{ci}^T\boldsymbol{\Gamma}_0\right]\right\}.$$

It follows that $L_{2,n}$ is maximized over $\boldsymbol{\Omega}_0$ by $\sum_{i=1}^{n}\boldsymbol{\Gamma}_0^T\mathbf{Y}_{ci}\mathbf{Y}_{ci}^T\boldsymbol{\Gamma}_0/n = \boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0$. Substituting back, we find the following partially maximized form for $L_{2,n}$:

$$L_{2,n}(\boldsymbol{\Gamma}_0|u) \;\simeq\; -(n/2)\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0|. \tag{4.9.5}$$

Holding $\boldsymbol{\Gamma}$ fixed, the log-likelihood $L_{1,n}$ is same as the log-likelihood for reduced rank regression of $\boldsymbol{\Gamma}^T\mathbf{Y}$ on $\mathbf{X}$. Therefore, by replacing $r \to u$, $\mathbf{Y} \to \boldsymbol{\Gamma}^T\mathbf{Y}$, $\mathbf{A} \to \boldsymbol{\eta}$, $\mathbf{B} \to \mathbf{B}$ and $\boldsymbol{\Sigma} \to \boldsymbol{\Omega}$ in (4.2.2) and in Lemma 4.1, we partially maximize $L_{1,n}(\boldsymbol{\Gamma}, \boldsymbol{\eta}, \mathbf{B}, \boldsymbol{\Omega}|d, u)$ over $\boldsymbol{\eta}$, $\mathbf{B}$ and $\boldsymbol{\Omega}$ and obtain the maximum likelihood estimators as

$$\begin{aligned}
\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}} &= \mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2}\mathbf{C}_{\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{S}_{\mathbf{X}}^{-1/2} \\
\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}} &= \mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2}\left\{\mathbf{I}_u - \mathbf{C}_{\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{C}_{\mathbf{X},\boldsymbol{\Gamma}^T\mathbf{Y}}^{(d)}\right\}\mathbf{S}_{\boldsymbol{\Gamma}^T\mathbf{Y}}^{1/2},
\end{aligned}$$

from which Lemma 4.2 follows.

Proof of Proposition 4.1 The log-likelihood function in (4.9.3) after partial maximization becomes

$$L_n(\boldsymbol{\Gamma}|d, u) \simeq -(n/2)\left\{\log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Gamma}_0| + \log|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}}|\right\}, \tag{4.9.6}$$

which lead us to the objective function $\mathrm{F}_n(\mathbf{G}|d, u) := (-2/n)L_n(\mathbf{G}|d, u)$ for numerical optimization over $\text{span}(\mathbf{G}) \in \mathcal{G}_{r,u}$. We next simplify the expression of $\log|\widehat{\boldsymbol{\Omega}}_{\mathbf{G}}|$ as

$$\begin{aligned}
\log|\widehat{\boldsymbol{\Omega}}_{\mathbf{G}}| &= \log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{1/2}\left\{\mathbf{I}_u - \mathbf{C}_{\mathbf{G}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{C}_{\mathbf{X},\mathbf{G}^T\mathbf{Y}}^{(d)}\right\}\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{1/2}| \\
&= 2\cdot\log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{1/2}| + \log|\mathbf{I}_u - \mathbf{C}_{\mathbf{G}^T\mathbf{Y},\mathbf{X}}^{(d)}\mathbf{C}_{\mathbf{X},\mathbf{G}^T\mathbf{Y}}^{(d)}| \\
&= \log|\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}| + \log|\mathbf{I}_u - \mathbf{S}_{\mathbf{Z}_{\mathbf{G}}\circ\mathbf{X}}^{(d)}|,
\end{aligned}$$

where $\mathbf{S}_{\mathbf{G}^T\mathbf{Y}} = \mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}^T$ and $\mathbf{Z}_{\mathbf{G}} = (\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y}$ is the standardized random vector in $\mathbb{R}^u$. Equation (4.3.2) is then obtained by noticing $\log|\mathbf{G}_0\mathbf{S}_{\mathbf{Y}}\mathbf{G}_0^T| = \log|\mathbf{S}_{\mathbf{Y}}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$ in the objective function (4.9.6).

We next prove the equality in (4.3.3). The first term in (4.3.3) can be re-expressed as $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}| = \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}| + \log|\mathbf{S}_{\mathbf{Z}_{\mathbf{G}}|\mathbf{X}}|$ according to the following.

$$\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G} = \mathbf{G}^T\mathbf{S}_{\mathbf{Y}\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{X}\mathbf{Y}}\mathbf{G} = \mathbf{S}_{\mathbf{G}^T\mathbf{Y},\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{X},\mathbf{G}^T\mathbf{Y}} = \mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Z}_{\mathbf{G}}|\mathbf{X}}\mathbf{S}_{\mathbf{G}^T\mathbf{Y}}^{1/2}.$$

The objective function in (4.3.3) now become

$$\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G}| + \log|\mathbf{S}_{\mathbf{Z_G}|\mathbf{X}}| + \log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}| + \sum_{i=d+1}^{u}\log[\widehat{\omega}_i(\mathbf{G})],$$

where $\widehat{\omega}_i(\mathbf{G})$ is the i-th eigenvalue of $\mathbf{S}_{\mathbf{Z_G}\circ\mathbf{X}}$. The equality connecting (4.3.2) and (4.3.3) is proved by noticing that $\mathbf{S}_{\mathbf{Z_G}|\mathbf{X}} = \mathbf{S}_{\mathbf{Z_G}} - \mathbf{S}_{\mathbf{Z_G}\circ\mathbf{X}} = \mathbf{I}_u - \mathbf{S}_{\mathbf{Z_G}\circ\mathbf{X}}$ and that the log-determinant of a positive definite matrix is the sum of the logarithms of its eigenvalues.

Proof of Proposition 4.3 The proof follows trivially by combining the results in Lemma 4.2 and Proposition 4.1.

4.9.2 Proposition 4.2

Recall that in (4.3.3), $\widehat{\omega}_i(\mathbf{G})$ is the i-th eigenvalues of the following matrix.

$$\begin{aligned}\mathbf{S}_{\mathbf{Z_G}|\mathbf{X}}^{-1} &= (\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2}(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}\mathbf{G})(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2} \\ &= \mathbf{I}_u + (\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2}(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}\circ\mathbf{X}}\mathbf{G})(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G})^{-1/2},\end{aligned}$$

which relies on the two sample covariance matrices: $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{Y}\circ\mathbf{X}}$. These two matrices are both positive semi-definite and converge to $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\mathbf{Y}\circ\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}$ with probability one as $n \to \infty$. Since $\text{rank}(\mathbf{\Sigma}_{\mathbf{Y}\circ\mathbf{X}}) = \text{rank}(\mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}) = \text{rank}(\boldsymbol{\beta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T) = d$, the last $(u-d)$ eigenvalues $\widehat{\omega}_j(\mathbf{G})$, $j = d+1, \ldots, u$, will equal to one with probability one as $n \to \infty$ for any value of $\mathbf{G}$. Therefore, as $n \to \infty$,

$$\sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\sum_{i=d+1}^{u}\log[\widehat{\omega}_i(\mathbf{G})]\right\} \xrightarrow{p} 0. \tag{4.9.7}$$

We next show that $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}|$ converges in probability to $\log|\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G}|$ uniformly in $\mathbf{G}$ by the following argument.

$$\begin{aligned}\delta(\mathbf{G}) &:= \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}| - \log|\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G}|\right\} \\ &= \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|(\mathbf{G}^T\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G})(\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G})^{-1}|\right\} \\ &= \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G})^{-1}\mathbf{G}^T|_0\right\} \\ &= \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\cdot\mathbf{\Sigma}^{-1/2}\mathbf{G}(\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1/2}|_0\right\} \\ &= \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{P}_{\mathbf{\Sigma}_{\mathbf{Y}}^{-1/2}\mathbf{G}}|_0\right\},\end{aligned}$$

where we use $|\cdot|_0$ to denote the product of the non-zero eigenvalues of a positive semi-definite matrix. We then can derive that

$$\delta(\mathbf{G}) = \sup_{\mathbf{G}\in\mathcal{G}_{r,u}}\left\{\log|\mathbf{P}_{\mathbf{\Sigma}_{\mathbf{Y}}^{-1/2}\mathbf{G}}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{P}_{\mathbf{\Sigma}_{\mathbf{Y}}^{-1/2}\mathbf{G}}|_0\right\}, \tag{4.9.8}$$

where $\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}$ was projected onto an u-dimensional subspace $\text{span}(\mathbf{\Sigma}_{\mathbf{Y}}^{-1/2}\mathbf{G})$. The quantity within $|\cdot|_0$ then has at most u nonzero eigenvalues. Because the projection matrix can not inflate the eigenvalues,

$$\delta(\mathbf{G}) \leq \sup_{\mathbf{G}\in\mathcal{G}_{r,u}} \left\{\log|\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{1/2}|_0\right\}, \tag{4.9.9}$$

which converges to zero in probability. Similarly, we can show that $\log|\mathbf{G}^T\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\mathbf{G}|$ converges in probability to $\log|\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}|$ uniformly in $\mathbf{G}$. Hence we have proved that the objective function $\text{F}_n(\mathbf{G}|d,u)$ in (4.3.3) converges in probability to $\text{F}(\mathbf{G}|u)$ uniformly in $\mathbf{G}$. The rest of the proof is similar to the proof of Proposition 4.2 in Cook et al. (2013) that

$$\begin{aligned}
\log|\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{G}| &= \log|\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{G}_0| \\
&= \log|\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T(\mathbf{\Sigma}+\boldsymbol{\beta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T)\mathbf{G}_0| \\
&\geq \log|\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}| + \log|\mathbf{G}_0^T\mathbf{\Sigma}\mathbf{G}_0| \\
&\geq \log|\mathbf{\Sigma}|,
\end{aligned}$$

where the first inequality achieves its lower bound if $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\mathbf{G})$; and the second inequality achieves its lower bound if $\text{span}(\mathbf{G})$ is a reducing subspace of $\mathbf{\Sigma}$. The uniqueness of the minimizer $\text{span}(\widehat{\mathbf{\Gamma}}) = \text{span}(\arg\min_{\mathbf{G}} \text{F}(\mathbf{G}|u))$ is guaranteed by the uniqueness of the envelope, which has dimension u.

4.9.3 Proposition 4.5

For notation convenience, we define two covariance matrices $\mathbf{M}_B := \mathbf{B}^T(\mathbf{B}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1}\mathbf{B} \leq \mathbf{\Sigma}_{\mathbf{X}}^{-1}$ and $\mathbf{M}_A := \mathbf{A}(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T \leq \mathbf{\Sigma}$. For any full row rank transformation $\mathbf{O} \in \mathbb{R}^{d\times q}$ we could replace $\mathbf{A}$ by $\mathbf{AO}$ and replace $\mathbf{B}$ by $\mathbf{OB}$ without changing the value of $\mathbf{M}_A$ or $\mathbf{M}_B$. Also the projection matrices $\mathbf{P}_{\mathbf{A}(\mathbf{\Sigma}^{-1})} = \mathbf{M}_A\mathbf{\Sigma}^{-1}$ and $\mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma}_{\mathbf{X}})} = \mathbf{M}_B\mathbf{\Sigma}_{\mathbf{X}}$.

Obtaining equation (4.4.3)

This result can be found in Anderson (1999) using canonical variables. We replicate the computation in our framework with details. Recall that the Fisher information is

$$\mathbf{J}_{\mathbf{h}} = \begin{pmatrix} \mathbf{J}_{\boldsymbol{\beta}} & 0 \\ 0 & \mathbf{J}_{\mathbf{\Sigma}} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{X}}\otimes\mathbf{\Sigma}^{-1} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_r^T(\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix}, \tag{4.9.10}$$

where $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \mathbf{J}_{\boldsymbol{\beta}}^{-1} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\otimes\mathbf{\Sigma}$.

By noticing $\mathbf{h}_1 = \text{vec}(\boldsymbol{\beta}) = \text{vec}(\mathbf{AB}) = (\mathbf{B}^T\otimes\mathbf{I}_r)\text{vec}(\mathbf{A}) = (\mathbf{I}_p\otimes\mathbf{A})\text{vec}(\mathbf{B})$, we have

$$\mathbf{H} = \begin{pmatrix} \mathbf{B}^T\otimes\mathbf{I}_r & \mathbf{I}_p\otimes\mathbf{A} & 0 \\ 0 & 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix} := \begin{pmatrix} \mathbf{H}_1 & 0 \\ 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix}. \tag{4.9.11}$$

Because of the similar block-diagonal structure in $\mathbf{J_h} = \text{diag}(\mathbf{J}_{\boldsymbol{\beta}}, \mathbf{J}_{\boldsymbol{\Sigma}})$, we can get

$$\mathbf{H}(\mathbf{H}^T\mathbf{J_h}\mathbf{H})^\dagger\mathbf{H}^T = \begin{pmatrix} \mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1)^\dagger\mathbf{H}_1^T & 0 \\ 0 & \mathbf{J}_{\boldsymbol{\Sigma}}^{-1} \end{pmatrix},$$

which means that $\boldsymbol{\beta} = \mathbf{AB}$ and $\boldsymbol{\Sigma}$ are orthogonal parameters in reduced-rank regression and the asymptotic covariance for $\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})$ is $\mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1)^\dagger\mathbf{H}_1^T$. Because $\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1$ is not full rank under the reduced rank regression model, we can not use the block-matrix inversion formula. However, notice that asymptotic covariance $\mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1)^\dagger\mathbf{H}_1^T$ depends only on the column space of $\mathbf{H}_1$, we thus could use any full row rank matrix $\mathbf{T}_1$ to get

$$\mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1)^\dagger\mathbf{H}_1^T = \mathbf{H}_1\mathbf{T}_1(\mathbf{T}_1^T\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1\mathbf{T}_1)^\dagger\mathbf{T}_1^T\mathbf{H}_1. \tag{4.9.12}$$

More specifically, we have each part

$$\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1 = \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1}\mathbf{A} \\ \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} \end{pmatrix} \tag{4.9.13}$$

$$\mathbf{T}_1 = \begin{pmatrix} \mathbf{I}_{rd} & -(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{A} \\ 0 & \mathbf{I}_{pd} \end{pmatrix}$$

$$\mathbf{H}_1\mathbf{T}_1 = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I}_r & (\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}) \otimes \mathbf{A} \end{pmatrix}. \tag{4.9.14}$$

where we have used $\mathbf{M}_B = \mathbf{B}^T(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1}\mathbf{B}$ for notation convenience. Then,

$$\mathbf{T}_1^T\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1\mathbf{T}_1 = \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1} & 0 \\ 0 & (\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}) \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} \end{pmatrix}.$$

To get the Moore-Penrose inverse of $\mathbf{T}_1^T\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1\mathbf{T}_1$, we first notice that it has rank $(p+r)d - d^2$ and the only non-invertable part is $(\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}})$ which causes rank deficiency of d^2. The Moore-Penrose inverse of $\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}$ is obtained as follows by noticing $\mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{M}_B = \mathbf{M}_B$.

$$(\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}})^\dagger = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B. \tag{4.9.15}$$

Therefore,

$$\left(\mathbf{T}_1^T\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1\mathbf{T}_1\right)^\dagger = \begin{pmatrix} (\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1} \otimes \boldsymbol{\Sigma} & 0 \\ 0 & (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} \end{pmatrix}. \tag{4.9.16}$$

The asymptotic covariance $\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] = \mathbf{H}_1\mathbf{T}_1(\mathbf{T}_1^T\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1\mathbf{T}_1)^{\dagger}\mathbf{T}_1^T\mathbf{H}_1$ is computed with (4.9.14),

$$\begin{aligned}
\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] &= \mathbf{M}_B \otimes \boldsymbol{\Sigma} \\
&+ (\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}})(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B)(\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}) \otimes \mathbf{M}_A \\
&= \mathbf{M}_B \otimes \boldsymbol{\Sigma} + (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \mathbf{M}_A \\
&= [\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B)] \otimes \boldsymbol{\Sigma} + (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \mathbf{M}_A \\
&= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} - (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes (\boldsymbol{\Sigma} - \mathbf{M}_A), \qquad (4.9.17)
\end{aligned}$$

then the equation (4.4.3) is derived from the following arguments.

$$\begin{aligned}
\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})] &= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} - (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes (\boldsymbol{\Sigma} - \mathbf{M}_A) \\
&= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} - \left[(\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\right] \otimes \left[(\mathbf{I}_r - \mathbf{M}_A\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}\right] \\
&= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} - \left[\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\right] \otimes \left[\mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}\right] \\
&= (\mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma})})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} \\
&= (\mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma})})\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})].
\end{aligned}$$

Obtaining equation (4.4.4)

The Fisher information for $(\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T)^T = [\text{vec}^T(\mathbf{A}), \text{vec}^T(\mathbf{B})]^T$ is given in (4.9.13) as

$$\mathbf{H}_1^T\mathbf{J}_{\boldsymbol{\beta}}\mathbf{H}_1 := \begin{pmatrix} \mathbf{J}_{\mathbf{A}} & \mathbf{J}_{\mathbf{AB}} \\ \mathbf{J}_{\mathbf{BA}} & \mathbf{J}_{\mathbf{B}} \end{pmatrix} = \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1}\mathbf{A} \\ \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T \otimes \mathbf{A}\boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} \end{pmatrix}. \qquad (4.9.18)$$

If we known $\mathbf{A}$, then we could cross the first row and the first column, and hence

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\mathbf{B}}_{\mathbf{A}})] = \mathbf{J}_{\mathbf{B}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}. \qquad (4.9.19)$$

Similarly,

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\mathbf{A}}_{\mathbf{B}})] = \mathbf{J}_{\mathbf{A}}^{-1} = (\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1} \otimes \boldsymbol{\Sigma}. \qquad (4.9.20)$$

Then by using the fact that $\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}}) = (\mathbf{I}_p \otimes \mathbf{A})\text{vec}(\widehat{\mathbf{B}}_{\mathbf{A}})$ and that $\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{B}}) = (\mathbf{B}^T \otimes \mathbf{I}_r)\text{vec}(\widehat{\mathbf{A}}_{\mathbf{B}})$, we have

$$\begin{aligned}
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}})] &= \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{M}_A \\
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{B}})] &= \mathbf{M}_B \otimes \boldsymbol{\Sigma}.
\end{aligned}$$

By noticing $\mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} = \mathbf{M}_A\boldsymbol{\Sigma}^{-1}$ and $\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} = \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}}$, we have

$$\begin{aligned}
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}}\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}^T)] &= [\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{I}_r](\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{M}_A)[\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}^T \otimes \mathbf{I}_r] \\
&= \left[(\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_{\mathbf{X}})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{X}}^T\mathbf{M}_B)\right] \otimes \mathbf{M}_A \\
&= (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \mathbf{M}_A \\
\text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\widehat{\boldsymbol{\beta}}_{\mathbf{B}})] &= [\mathbf{I}_p \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}](\mathbf{M}_B \otimes \boldsymbol{\Sigma})[\mathbf{I}_p \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}^T] \\
&= \mathbf{M}_B \otimes (\mathbf{I}_r - \mathbf{M}_A\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}(\mathbf{I}_r - \boldsymbol{\Sigma}^{-1}\mathbf{M}_A) \\
&= \mathbf{M}_B \otimes (\boldsymbol{\Sigma} - \mathbf{M}_A).
\end{aligned}$$

The proof of Proposition 4.5 is then completed by compare the above quantities with (4.9.17).

4.9.4 Proposition 4.6

The role of $\boldsymbol{\eta}$ is analogous to $\mathbf{A}$ given $\boldsymbol{\Gamma}$, thus we define $\mathbf{M}_\eta := \boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T \leq \boldsymbol{\Omega}$. Note that the projection matrices $\mathbf{P}_{\boldsymbol{\eta}(\boldsymbol{\Omega}^{-1})} = \mathbf{M}_\eta\boldsymbol{\Omega}^{-1}$.

Explicit expression for the asymptotic covariance

We compute the explicit expression for $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$ in this section. By noticing $\mathbf{h}_1 = \text{vec}(\boldsymbol{\beta}) = \text{vec}(\boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{B}) = (\boldsymbol{\eta}^T\mathbf{B}^T \otimes \mathbf{I}_r)\text{vec}(\boldsymbol{\Gamma}) = (\mathbf{B}^T \otimes \boldsymbol{\Gamma})\text{vec}(\boldsymbol{\eta}) = (\mathbf{I}_p \otimes \boldsymbol{\Gamma}\boldsymbol{\eta})\text{vec}(\mathbf{B})$, we have

$$\mathbf{R} = \begin{pmatrix} \mathbf{B}^T\boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{B}^T \otimes \boldsymbol{\Gamma} & \mathbf{I}_p \otimes \boldsymbol{\Gamma}\boldsymbol{\eta} & 0 & 0 \\ 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & 0 & 0 & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} \end{pmatrix}. \tag{4.9.21}$$

The asymptotic covariance $\text{avar}(\sqrt{n}\mathbf{h}(\widehat{\boldsymbol{\phi}})) = \mathbf{R}(\mathbf{R}^T\mathbf{J}_\mathbf{h}\mathbf{R})^\dagger\mathbf{R}^T = \widetilde{\mathbf{R}}(\widetilde{\mathbf{R}}^T\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}})^\dagger\widetilde{\mathbf{R}}^T$ for any $\widetilde{\mathbf{R}}$ such that $\mathbf{R} = \widetilde{\mathbf{R}}\mathbf{T}$ for a full row rank matrix $\mathbf{T}$. We choose $\widetilde{\mathbf{R}}$ to make $\widetilde{\mathbf{R}}^T\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}}$ block-diagonal as follows.

$$\widetilde{\mathbf{R}} = \begin{pmatrix} \mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & \mathbf{B}^T \otimes \boldsymbol{\Gamma} & (\mathbf{I}_p - \mathbf{M}_B\boldsymbol{\Sigma}_\mathbf{X}) \otimes \boldsymbol{\Gamma}\boldsymbol{\eta} & 0 & 0 \\ 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & 0 & 0 & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} \end{pmatrix}, \tag{4.9.22}$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{I}_u \otimes \boldsymbol{\Gamma}_0^T & 0 & 0 & 0 & 0 \\ \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}^T & \mathbf{I}_{ud} & (\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T)^{-1}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X} \otimes \boldsymbol{\eta} & 0 & 0 \\ 0 & 0 & \mathbf{I}_{pd} & 0 & 0 \\ 2\mathbf{C}_u(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}^T) & 0 & 0 & \mathbf{I}_{\frac{1}{2}r(r+1)} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I}_{\frac{1}{2}(r-u)(r-u+1)} \end{pmatrix}. \tag{4.9.23}$$

Next, we calculate $\widetilde{\mathbf{R}}^T\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}}$ and verify that it is block-diagonal. We decompose $\widetilde{\mathbf{G}}$ by it 2×5 blocks as $\widetilde{\mathbf{R}} := (\widetilde{\mathbf{G}}_1, \widetilde{\mathbf{G}}_2, \widetilde{\mathbf{G}}_3, \widetilde{\mathbf{G}}_4, \widetilde{\mathbf{G}}_5)$. We first calculate $\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}}$ and write down the 2×5 blocks by column:

$$\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}}_1 = \begin{pmatrix} \boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} \\ \mathbf{E}_r^T(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0) \end{pmatrix}, \tag{4.9.24}$$

$$\mathbf{J}_\mathbf{h}[\widetilde{\mathbf{R}}_2, \widetilde{\mathbf{R}}_3] = \begin{pmatrix} \boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} & (\boldsymbol{\Sigma}_\mathbf{X} - \boldsymbol{\Sigma}_\mathbf{X}\mathbf{M}_B\boldsymbol{\Sigma}_\mathbf{X}) \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \\ 0 & 0 \end{pmatrix}, \tag{4.9.25}$$

$$\mathbf{J}_\mathbf{h}[\widetilde{\mathbf{G}}_4, \widetilde{\mathbf{G}}_5] = \begin{pmatrix} 0 & 0 \\ \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1})\mathbf{E}_u & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1})\mathbf{E}_{r-u} \end{pmatrix}. \tag{4.9.26}$$

Then $\widetilde{\mathbf{R}}^T\mathbf{J_h}\widetilde{\mathbf{R}}$ equals to a block-diagonal matrix with five blocks: $\widetilde{\mathbf{R}}_i^T\mathbf{J_h}\widetilde{\mathbf{R}}_i$, $i=1,\ldots,5$. The explicit expressions are given as follows.

$$\begin{aligned}
\widetilde{\mathbf{R}}_1^T\mathbf{J_h}\widetilde{\mathbf{R}}_1 &= \boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T\otimes\boldsymbol{\Omega}_0^{-1} \\
&\quad +2(\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\mathbf{C}_r^T\mathbf{E}_r^T(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}_0) \\
&= \boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T\otimes\boldsymbol{\Omega}_0^{-1}+\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-2\mathbf{I}_{u(r-u)}+\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0, \\
\widetilde{\mathbf{R}}_2^T\mathbf{J_h}\widetilde{\mathbf{R}}_2 &= \mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\otimes\boldsymbol{\Omega}^{-1}, \\
\widetilde{\mathbf{R}}_3^T\mathbf{J_h}\widetilde{\mathbf{R}}_3 &= \boldsymbol{\Sigma}_\mathbf{X}\otimes\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}, \\
\widetilde{\mathbf{R}}_4^T\mathbf{J_h}\widetilde{\mathbf{R}}_4 &= \mathbf{E}_u^T(\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}^T)\mathbf{C}_r^T\cdot\frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1})\mathbf{E}_u \\
&= \frac{1}{2}\mathbf{E}_u^T(\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}^T)(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1})\mathbf{E}_u \\
&= \frac{1}{2}\mathbf{E}_u^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_u.
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathbf{R}}_5^T\mathbf{J_h}\widetilde{\mathbf{R}}_5 &= \mathbf{E}_{r-u}^T(\boldsymbol{\Gamma}_0^T\otimes\boldsymbol{\Gamma}_0^T)\mathbf{C}_r^T\cdot\frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1})\mathbf{E}_{r-u} \\
&= \frac{1}{2}\mathbf{E}_{r-u}^T(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathbf{E}_{r-u}
\end{aligned}$$

Then the asymptotic covariance is

$$\text{avar}[\sqrt{n}\mathbf{h}(\widehat{\boldsymbol{\phi}})] = \sum_{i=1}^{5}\widetilde{\mathbf{R}}_i(\widetilde{\mathbf{R}}_i^T\mathbf{J_h}\widetilde{\mathbf{R}}_i)^\dagger\widetilde{\mathbf{R}}_i^T$$

We are only interested in the asymptotic covariance of $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$, which is the upper left block of $\text{avar}[\sqrt{n}\mathbf{h}(\widehat{\boldsymbol{\phi}})]$. And $\widetilde{\mathbf{R}}_4(\widetilde{\mathbf{R}}_4^T\mathbf{J_h}\widetilde{\mathbf{R}}_4)^\dagger\widetilde{\mathbf{R}}_4^T$ and $\widetilde{\mathbf{R}}_5(\widetilde{\mathbf{R}}_5^T\mathbf{J_h}\widetilde{\mathbf{R}}_5)^\dagger\widetilde{\mathbf{R}}_5^T$ have no contribution to that. So we will focus our attention on the upper left block of $\widetilde{\mathbf{R}}_i(\widetilde{\mathbf{R}}_i^T\mathbf{J_h}\widetilde{\mathbf{R}}_i)^\dagger\widetilde{\mathbf{R}}_i^T$, $i=1,2,3$. The upper left block of $\widetilde{\mathbf{R}}_1(\widetilde{\mathbf{R}}_1^T\mathbf{J_h}\widetilde{\mathbf{R}}_1)^\dagger\widetilde{\mathbf{R}}_1^T$ is

$$\begin{aligned}
&(\mathbf{B}^T\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)(\widetilde{\mathbf{R}}_1^T\mathbf{J_h}\widetilde{\mathbf{R}}_1)^\dagger(\boldsymbol{\eta}\mathbf{B}\otimes\boldsymbol{\Gamma}_0^T) \\
= \;&(\mathbf{B}^T\boldsymbol{\eta}^T\otimes\boldsymbol{\Gamma}_0)(\boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T\otimes\boldsymbol{\Omega}_0^{-1}+\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-2\mathbf{I}_{u(r-u)}+\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0)^\dagger(\boldsymbol{\eta}\mathbf{B}\otimes\boldsymbol{\Gamma}_0^T).
\end{aligned}$$

The upper left block of $\widetilde{\mathbf{R}}_2(\widetilde{\mathbf{R}}_2^T\mathbf{J_h}\widetilde{\mathbf{R}}_2)^\dagger\widetilde{\mathbf{R}}_2^T$ is

$$\begin{aligned}
&(\mathbf{B}^T\otimes\boldsymbol{\Gamma})(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\otimes\boldsymbol{\Omega}^{-1})^\dagger(\mathbf{B}\otimes\boldsymbol{\Gamma}^T) \\
= \;&(\mathbf{B}^T\otimes\boldsymbol{\Gamma})[(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T)^{-1}\otimes\boldsymbol{\Omega}](\mathbf{B}\otimes\boldsymbol{\Gamma}^T) \\
= \;&\mathbf{M}_B\otimes\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T.
\end{aligned}$$

The upper left block of $\widetilde{\mathbf{R}}_3(\widetilde{\mathbf{R}}_3^T\mathbf{J_h}\widetilde{\mathbf{R}}_3)^\dagger\widetilde{\mathbf{R}}_3^T$ is

$$\begin{aligned}
&[(\mathbf{I}_p-\mathbf{M}_B\boldsymbol{\Sigma}_\mathbf{X})\otimes\boldsymbol{\Gamma}\boldsymbol{\eta}](\boldsymbol{\Sigma}_\mathbf{X}\otimes\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^\dagger[(\mathbf{I}_p-\boldsymbol{\Sigma}_\mathbf{X}\mathbf{M}_B)\otimes\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T] \\
= \;&(\boldsymbol{\Sigma}_\mathbf{X}^{-1}-\mathbf{M}_B)\otimes\boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T,
\end{aligned}$$

where $\mathbf{M}_\eta = \boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T$.

Hence, the asymptotic covariance avar$[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})]$ equals to

$$\begin{aligned} & (\mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)(\boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} - 2\mathbf{I}_{u(r-u)} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0)^\dagger(\boldsymbol{\eta}\mathbf{B} \otimes \boldsymbol{\Gamma}_0^T) \\ + \quad & \mathbf{M}_B \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\Sigma}_\mathbf{X}^{-1} - \mathbf{M}_B) \otimes \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T. \end{aligned} \tag{4.9.27}$$

Interpretation

The Fisher information matrix for $\widehat{\boldsymbol{\phi}}$ is simply $\mathbf{R}^T\mathbf{J}_\mathbf{h}\mathbf{R}$:

$$\mathbf{R}^T\mathbf{J}\mathbf{R} := \begin{pmatrix} \mathbf{J}_{\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\Gamma}\mathbf{B}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}} & 0 \\ \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\eta}\mathbf{B}} & 0 & 0 \\ \mathbf{J}_{\mathbf{B}\boldsymbol{\Gamma}} & \mathbf{J}_{\mathbf{B}\boldsymbol{\eta}} & \mathbf{J}_{\mathbf{B}} & 0 & 0 \\ \mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}} & 0 & 0 & \mathbf{J}_{\boldsymbol{\Omega}} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{J}_{\boldsymbol{\Omega}_0} \end{pmatrix}. \tag{4.9.28}$$

Each nonzero block is

$$\begin{aligned} \mathbf{J}_{\boldsymbol{\Gamma}} &= \boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Gamma})\mathbf{K}_{ru} \\ &\quad + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T - 2\mathbf{I}_u \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T \\ \mathbf{J}_{\boldsymbol{\eta}} &= \mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T \otimes \boldsymbol{\Omega}^{-1} \\ \mathbf{J}_{\mathbf{B}} &= \boldsymbol{\Sigma}_\mathbf{X} \otimes \boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \\ \mathbf{J}_{\boldsymbol{\Omega}} &= \frac{1}{2}\mathbf{E}_u^T(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_u \\ \mathbf{J}_{\boldsymbol{\Omega}_0} &= \frac{1}{2}\mathbf{E}_{r-u}^T(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\mathbf{E}_{r-u}. \\ \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\eta}} &= \boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \\ \mathbf{J}_{\boldsymbol{\Gamma}\mathbf{B}} &= \boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \\ \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}} &= (\mathbf{I}_u \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1})\mathbf{E}_u \\ \mathbf{J}_{\boldsymbol{\eta}\mathbf{B}} &= \mathbf{B}\boldsymbol{\Sigma}_\mathbf{X} \otimes \boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \end{aligned}$$

Asymptotic covariance when $\boldsymbol{\eta}$ and $\mathbf{B}$ are known

The asymptotic covariance for vec$(\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta},\mathbf{B}})$ is

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta},\mathbf{B}})] = \left(\mathbf{J}_{\boldsymbol{\Gamma}} - \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}}\right)^{-1}.$$

Follow Cook et al. (2010), we have

$$\begin{aligned} \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta},\mathbf{B}})] &= [\boldsymbol{\eta}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{B}^T\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T - 2\mathbf{I}_u \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T \\ &\quad + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T]^\dagger, \end{aligned}$$

and by replacing $\boldsymbol{\eta}\mathbf{B} \to \boldsymbol{\eta}$ in Cook et al. (2010), it is easy to obtain the following results

$$\text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})] = \left[\widetilde{\mathbf{R}}_1(\widetilde{\mathbf{R}}_1^T\mathbf{J}_\mathbf{h}\widetilde{\mathbf{R}}_1)^\dagger\widetilde{\mathbf{R}}_1^T\right]_{11},$$

where $[]_{11}$ means the upper left block of a block-wise matrix. The above equality explains the contribution from the first column block of $\widetilde{\mathbf{R}}$, which is the first term in (4.9.27).

Therefore, (4.9.27) can be written as

$$\begin{aligned} \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RE}})] &= \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})] \\ &\quad +\mathbf{M}_B \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T \qquad (4.9.29) \\ &= \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{B}})] + \text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})], \qquad (4.9.30) \end{aligned}$$

where the last equality follows from the asymptotic covariance of $\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{RR}})$ in (4.9.17) and from Lemma 4.1 that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ is $\boldsymbol{\Gamma}$ times the reduced-rank regression estimator of regression $\boldsymbol{\Gamma}^T\mathbf{Y}$ on $\mathbf{X}$.

Asymptotic covariance when $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are known

The asymptotic covariance for $\text{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma},\mathbf{B}})$ is

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma},\mathbf{B}})] = \mathbf{J}_{\boldsymbol{\eta}}^{-1} = (\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{B}^T)^{-1} \otimes \boldsymbol{\Omega}. \qquad (4.9.31)$$

Notice that $\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\mathbf{B}}) = \text{vec}(\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma},\mathbf{B}}\mathbf{B}) = (\mathbf{B}^T \otimes \boldsymbol{\Gamma})\text{vec}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma},\mathbf{B}})$, we have

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\mathbf{B}})] = \mathbf{M}_B \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T. \qquad (4.9.32)$$

Asymptotic covariance when $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}$ are known

The asymptotic covariance for $\text{vec}(\widehat{\mathbf{B}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}})$ is

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\mathbf{B}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}})] = \mathbf{J}_{\mathbf{B}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})^{-1}. \qquad (4.9.33)$$

Notice that $\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}}) = \text{vec}(\boldsymbol{\Gamma}\boldsymbol{\eta}\widehat{\mathbf{B}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}}) = (\mathbf{I}_p \otimes \boldsymbol{\Gamma}\boldsymbol{\eta})\text{vec}(\widehat{\mathbf{B}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}})$, we have

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}})] = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T. \qquad (4.9.34)$$

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma},\boldsymbol{\eta}}\mathbf{Q}^T_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})})] = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T. \qquad (4.9.35)$$

Decomposition

Finally, plugging (4.9.32) and (4.9.35) into (4.9.29), we have proven this Proposition.

4.9.5 Corollary 4.1

By noticing $\mathbf{A} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, we can write

$$\begin{aligned} \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} &= \boldsymbol{\Gamma}\boldsymbol{\eta}(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\mathbf{P}_{\boldsymbol{\eta}(\boldsymbol{\Omega}^{-1})}\boldsymbol{\Gamma}^T. \\ \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T &= \boldsymbol{\Gamma}\mathbf{P}_{\boldsymbol{\eta}(\boldsymbol{\Omega}^{-1})}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \boldsymbol{\Gamma}\mathbf{P}_{\boldsymbol{\eta}(\boldsymbol{\Omega}^{-1})}\boldsymbol{\Gamma}^T \cdot \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma} = \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}. \end{aligned}$$

Then, from (4.9.29), we have

$$\begin{aligned}
\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] &= \mathbf{M}_B \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{M}_B) \otimes \boldsymbol{\Gamma}\mathbf{M}_\eta\boldsymbol{\Gamma}^T \\
&= \mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma} + \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma} \\
&= \left\{\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{I}_r + \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\right\}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma} \\
&= \left\{\mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\right\}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma} \\
&= \left\{\mathbf{I}_{pr} - \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \otimes \mathbf{Q}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\right\}(\mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}})\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}.
\end{aligned}$$

4.9.6 Proposition 4.7

From Proposition 4.1 and Proposition 4.3, we see that the minimizer $\widehat{\mathbf{h}}_{\text{RE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ of $\mathcal{F}(\mathbf{h}(\boldsymbol{\phi}), \widehat{\mathbf{h}}_{\text{OLS}})$ is Fisher consistent. The rest of the proof relies on Shapiro's (1986) results on the asymptotics of overparameterized structural models. In order to apply Shapiro's (1986) theory in our context, we first can check that $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}})$ satisfies: (1) $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}}) \geq 0$ for all $\widehat{\mathbf{h}}_{\text{OLS}}$ and $\mathbf{h}$; (2) $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}}) = 0$ if and only if $\widehat{\mathbf{h}}_{\text{OLS}} = \mathbf{h}$; and (3) $\mathcal{F}(\mathbf{h}, \widehat{\mathbf{h}}_{\text{OLS}})$ is twice continuously differentiable in $\mathbf{h}$ and $\widehat{\mathbf{h}}_{\text{OLS}}$. Recall from Section 4.4.2 that we use the subscript 0 to emphasize the true parameter: $\mathbf{h}_0$ and $\boldsymbol{\phi}_0$ correspond to the true distribution of $\boldsymbol{\epsilon}$. Then $\widehat{\boldsymbol{h}}_{\text{OLS}}$ is $\sqrt{n}$-consistent for $\mathbf{h}_0$. Notice that $\widehat{\mathbf{h}}_{\text{OLS}}$ is a smooth function of the sample covariance matrices which converges in distribution to the population covariance matrices, then by the delta method we know $\sqrt{n}(\widehat{\mathbf{h}}_{\text{OLS}} - \mathbf{h}_0) \to N(0, \mathbf{K})$, for some positive definite covariance $\mathbf{K}$. Using Shapiro's (1986) Proposition 3.1 and Proposition 4.1, we will have $\sqrt{n}$-consistency results for $\widehat{\mathbf{h}}_{\text{RE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ as shown in Proposition 4.7.

4.9.7 Some technical derivations for Section 4.8

Following Cook et al. (2013), we define $\mathbf{C} = (\mathbf{X}^T, \mathbf{Y}^T)^T \in \mathbb{R}^{p+r}$. The likelihood based objective function for the multivariate linear model is

$$\text{F}(\boldsymbol{\Sigma}_{\mathbf{C}}) = \log|\boldsymbol{\Sigma}_{\mathbf{C}}| + \text{trace}(\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\mathbf{S}_{\mathbf{C}}), \tag{4.9.36}$$

where $\boldsymbol{\Sigma}_{\mathbf{C}}$ has one-to-one relationship with our parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathbf{X}}, \boldsymbol{\Sigma})$. From the parameterization of $\boldsymbol{\beta} = \mathbf{A}\mathbf{b}\boldsymbol{\Gamma}^T$ and $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, we have the following

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{C}} &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T \\ \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma} + \boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ \boldsymbol{\beta} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & 0 \\ 0 & \boldsymbol{\Sigma} \end{pmatrix}\begin{pmatrix} \mathbf{I} & \boldsymbol{\beta}^T \\ 0 & \mathbf{I} \end{pmatrix}, \\
\boldsymbol{\Sigma}_{\mathbf{C}}^{-1} &= \begin{pmatrix} \mathbf{I} & -\boldsymbol{\beta}^T \\ 0 & \mathbf{I} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{I} & 0 \\ -\boldsymbol{\beta} & \mathbf{I} \end{pmatrix}.
\end{aligned}$$

Therefore,

$$\log|\boldsymbol{\Sigma}_{\mathbf{C}}| = \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\boldsymbol{\Sigma}| = \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \log|\boldsymbol{\Sigma}|, \tag{4.9.37}$$

$$\begin{aligned}
\text{trace}(\mathbf{\Sigma_C^{-1}S_C}) &= \text{trace}\left\{\begin{pmatrix} \mathbf{\Sigma_X^{-1}} & 0 \\ 0 & \mathbf{\Sigma^{-1}} \end{pmatrix}\begin{pmatrix} \mathbf{I} & 0 \\ -\boldsymbol{\beta} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{S_X} & \mathbf{S_{XY}} \\ \mathbf{S_{YX}} & \mathbf{S_Y} \end{pmatrix}\begin{pmatrix} \mathbf{I} & -\boldsymbol{\beta}^T \\ 0 & \mathbf{I} \end{pmatrix}\right\} \\
&= \text{trace}(\mathbf{\Sigma_X^{-1}S_X}) + \text{trace}[\mathbf{\Sigma}^{-1}(\mathbf{S_Y} - \boldsymbol{\beta}\mathbf{S_{XY}} - \mathbf{S_{YX}}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{S_X}\boldsymbol{\beta}^T)] \\
&= \text{trace}(\mathbf{\Omega}^{-1}\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma}) + \text{trace}(\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^T\mathbf{S_X}\mathbf{\Gamma}_0) \\
&\quad +\text{trace}[\mathbf{\Sigma}^{-1}(\mathbf{S_Y} - \boldsymbol{\beta}\mathbf{S_{XY}} - \mathbf{S_{YX}}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{S_X}\boldsymbol{\beta}^T)].
\end{aligned}$$

Hence, it is easy to partially minimized the objective function $\log|\mathbf{\Sigma_C}| + \text{trace}(\mathbf{\Sigma_C^{-1}S_C})$ over $\mathbf{\Omega}$, $\mathbf{\Omega}_0$ and $\mathbf{\Sigma}$ that

$$\begin{aligned}
\widehat{\mathbf{\Omega}}_{\mathbf{\Gamma}} &= \mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma}, \\
\widehat{\mathbf{\Omega}}_{0,\mathbf{\Gamma}} &= \mathbf{\Gamma}_0^T\mathbf{S_X}\mathbf{\Gamma}_0, \\
\widehat{\mathbf{\Sigma}}_{\boldsymbol{\beta}} &= \mathbf{S_Y} - \boldsymbol{\beta}\mathbf{S_{XY}} - \mathbf{S_{YX}}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{S_X}\boldsymbol{\beta}^T.
\end{aligned}$$

Then the rest parameters and the partially minimized objective function are

$$\begin{aligned}
\text{F}(\mathbf{A}, \mathbf{b}, \mathbf{\Gamma}) &= \log|\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}_0^T\mathbf{S_X}\mathbf{\Gamma}_0| \\
&\quad + \log|\mathbf{S_Y} - \mathbf{Ab}\mathbf{\Gamma}^T\mathbf{S_{XY}} - \mathbf{S_{YX}}\mathbf{\Gamma}\mathbf{b}^T\mathbf{A}^T + \mathbf{Ab}\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma}\mathbf{b}^T\mathbf{A}^T| \qquad (4.9.38)
\end{aligned}$$

where the last term can be derived by maximizing the log-likelihood of reduced-rank regression of $\mathbf{Y}$ on $\mathbf{\Gamma}^T\mathbf{X}$. Hence we can minimize $\text{F}(\mathbf{A}, \mathbf{b}, \mathbf{\Gamma})$ over $\mathbf{A}$ and $\mathbf{b}$ to get

$$\begin{aligned}
\widehat{\mathbf{A}}_{\mathbf{\Gamma}} &= \mathbf{S_{YX}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T, \\
\widehat{\mathbf{b}}_{\mathbf{\Gamma}} &= \mathbf{V}_d^T(\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma}, \mathbf{\Gamma}^T\mathbf{S_{X\circ Y}}\mathbf{\Gamma}).
\end{aligned}$$

Then the last term of $\text{F}(\mathbf{A}, \mathbf{b}, \mathbf{\Gamma})$ in (4.9.38) become a function of $\mathbf{\Gamma}$:

$$\begin{aligned}
& \log|\mathbf{S_Y} - \mathbf{S_{YX}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T\widehat{\mathbf{b}}_{\mathbf{\Gamma}}\mathbf{\Gamma}^T\mathbf{S_{XY}}| \\
\simeq\ & \log|\mathbf{I}_r - \mathbf{S_Y^{-1/2}}\mathbf{S_{YX}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T\widehat{\mathbf{b}}_{\mathbf{\Gamma}}\mathbf{\Gamma}^T\mathbf{S_{XY}}\mathbf{S_Y^{-1/2}}| \\
=\ & \log|\mathbf{I}_d - \widehat{\mathbf{b}}_{\mathbf{\Gamma}}\mathbf{\Gamma}^T\mathbf{S_{XY}}\mathbf{S_Y^{-1/2}}\mathbf{S_Y^{-1/2}}\mathbf{S_{YX}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T| \\
=\ & \log|\mathbf{I}_d - \widehat{\mathbf{b}}_{\mathbf{\Gamma}}\mathbf{\Gamma}^T\mathbf{S_{X\circ Y}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T| \\
=\ & \log|\mathbf{I}_d - \widehat{\mathbf{b}}_{\mathbf{\Gamma}}\mathbf{\Gamma}^T\mathbf{S_{X\circ Y}}\mathbf{\Gamma}\widehat{\mathbf{b}}_{\mathbf{\Gamma}}^T| \\
=\ & \sum_{i=1}^{d}\log[1 - \widehat{\nu}_i(\mathbf{\Gamma})],
\end{aligned}$$

where $\widehat{\nu}_i(\mathbf{\Gamma})$ is the i-th largest eigenvalue of

$$(\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma})^{-1/2}\mathbf{\Gamma}^T\mathbf{S_{X\circ Y}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma})^{-1/2}, \qquad (4.9.39)$$

and by noticing that $\mathbf{S_{X\circ Y}} = \mathbf{S_X} - \mathbf{S_{X|Y}}$, we see that $1 - \widehat{\nu}_i(\mathbf{\Gamma})$ is the i-th smallest eigenvalue of

$$(\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma})^{-1/2}\mathbf{\Gamma}^T\mathbf{S_{X|Y}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S_X}\mathbf{\Gamma})^{-1/2}, \qquad (4.9.40)$$

and $\widehat{\omega}_i(\mathbf{\Gamma}) = 1/[1 - \nu_i(\mathbf{\Gamma})]$ is the i-th largest eigenvalue of

$$(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \mathbf{\Gamma})^{-1/2} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \mathbf{\Gamma})^{-1/2}. \tag{4.9.41}$$

Hence we can write the objective function for $\mathbf{\Gamma}$ as

$$\mathrm{F}_n(\mathbf{\Gamma}|d, u) := \log|\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma}| + \log|\mathbf{\Gamma}_0^T \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma}_0| - \sum_{i=1}^{d} \log|\widehat{\omega}_i(\mathbf{\Gamma})|. \tag{4.9.42}$$

And follow the same derivation in Section 4.3.2, we have

$$\mathrm{F}_n(\mathbf{\Gamma}|d, u) = \log|\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{\Gamma}| + \sum_{i=d+1}^{u} \log|\widehat{\omega}_i(\mathbf{\Gamma})|. \tag{4.9.43}$$

Chapter 5

Foundations for Envelope Models and Methods

5.1 Introduction

Envelope applications have so far been mostly restricted to the homoscedastic multivariate linear model

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{5.1.1}$$

where $\mathbf{Y} \in \mathbb{R}^r$, the predictor vector $\mathbf{X} \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, $\boldsymbol{\alpha} \in \mathbb{R}^r$ and the errors $\boldsymbol{\varepsilon}_i$ are independent copies of the normal random vector $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})$. As we reviewed in Section 2.2.2, given the dimension of the envelope, the envelope estimators in the three articles (Cook et al. 2010, 2013; Su and Cook 2011) are all maximum likelihood estimators based on normality assumptions.

In Cook et al. (2010), the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ of $\boldsymbol{\beta}$ was obtained by parameterizing model (5.1.1) in terms of a basis $\boldsymbol{\Gamma}$ for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ and then using maximum likelihood estimation, resulting in $\widehat{\boldsymbol{\beta}}_{\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\Gamma}}$ is any basis for the estimated envelope. They proved that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is never larger than that for $\widehat{\boldsymbol{\beta}}$, which happens because the envelope estimator accounts for the immaterial information $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ in the response prior to estimation. The reduction in variation achieved by the envelope estimator can be substantial when the immaterial variation $\text{var}(\boldsymbol{\Gamma}_0^T\mathbf{Y})$ is large relative to the material variation $\text{var}(\boldsymbol{\Gamma}^T\mathbf{Y})$.

Su and Cook (2011) used the $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$-envelope of $\text{span}(\boldsymbol{\beta}_1)$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, to develop a partial envelope estimator of $\boldsymbol{\beta}_1$ in the partitioned multivariate linear regression

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_{1i} + \boldsymbol{\beta}_2\mathbf{X}_{2i} + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n, \tag{5.1.2}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^r$ is the parameter vector of interest, $\mathbf{X} = (X_1, \mathbf{X}_2^T)^T$ and the remaining terms are as defined for model (5.1.1). In this formulation, the immaterial information is

$\boldsymbol{\Gamma}_0^T\mathbf{Y}$, where $\boldsymbol{\Gamma}_0$ is a basis for $\mathcal{E}^{\perp}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$. Since $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, the partial envelope estimator $\widehat{\boldsymbol{\beta}}_{1,\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\beta}}_1$ has the potential to yield efficiency gains beyond those for the full envelope, particularly when $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathbb{R}^r$ so the full envelope offers no gain. Su and Cook (2011) also gave an intuitive introduction to envelopes in the context of model (5.1.1).

Cook, et al. (2013) studied predictor reduction in model (5.1.1), except the predictors are now stochastic with $\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}}$. Their reasoning, which paralleled that of Cook, et al. (2010), lead them to parameterize the linear model in terms of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and to achieve similar substantial gains in the estimation of $\boldsymbol{\beta}$ and in prediction. The immaterial information in this setting is given by $\boldsymbol{\Gamma}_0^T\mathbf{X}$, where $\boldsymbol{\Gamma}_0$ is now a basis for $\mathcal{E}^{\perp}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$. They also showed that the SIMPLS algorithm for partial least squares provides a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and demonstrated that the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\beta}}\mathbf{P}^T_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}})}$ typically outperforms the SIMPLS estimator in practice.

It can be shown that the partial maximized log-likelihood functions $L_n(\boldsymbol{\Gamma}) = -(n/2)J_n(\boldsymbol{\Gamma})$ for estimation of a basis $\boldsymbol{\Gamma}$ of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$ or $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ all have the same form with

$$J_n(\boldsymbol{\Gamma}) = \log|\boldsymbol{\Gamma}^T\widehat{\mathbf{M}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\boldsymbol{\Gamma}|, \tag{5.1.3}$$

where the positive definite matrix $\widehat{\mathbf{M}}$ and the positive semi-definite matrix $\widehat{\mathbf{U}}$ depend on context. An estimated basis is then $\widehat{\boldsymbol{\Gamma}} = \arg\min J_n(\boldsymbol{\Gamma})$, where the minimization is carried out over a set of semi-orthogonal matrices whose dimensions depend on the envelope being estimated. Estimates of the remaining parameters are then simple functions of $\widehat{\boldsymbol{\Gamma}}$. To determine the estimators of the response envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, the predictor envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T)$ and the partial envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, we have $\{\widehat{\mathbf{M}}, \widehat{\mathbf{M}}+\widehat{\mathbf{U}}\} = \{\mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \mathbf{S}_{\mathbf{Y}}\}$, $\{\mathbf{S}_{\mathbf{X}|\mathbf{Y}}, \mathbf{S}_{\mathbf{X}}\}$ and $\{\mathbf{S}_{\mathbf{Y}|\mathbf{X}}, \mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}\}$, respectively. Techniques for estimating the dimension of an envelope are discussed in the parent articles of these methods, including use of an information criterion like AIC and BIC, cross validation or a hold-out sample.

The previous work on envelope models and methods is limited to multivariate linear regression. While envelope constructions seem natural and intuitive in that setting, nothing is available to guide the construction of envelopes in other contexts like generalized linear models. In this chapter, we introduce the constructive principle that an asymptotically normal estimator $\widehat{\boldsymbol{\phi}}$ of a parameter vector $\boldsymbol{\phi}$ may be improved by enveloping $\text{span}(\boldsymbol{\phi})$ with respect to the asymptotic covariance of $\widehat{\boldsymbol{\phi}}$. This principle recovers past estimators and allows for extensions to many other contexts. The 1D algorithm in Chapter 2 together with our proposed general framework provides $\sqrt{n}$-consistent envelope estimators under fairly weak moment conditions.

The rest of this chapter is organized as follows. In Section 5.2 we describe a general envelope construction that subsumes the known methods reviewed above. The simulation in Section 5.2.3 illustrates the advantages of this general envelope construction applied

to logistic regression with correlated predictors. We turn to envelopes in likelihood-based estimation in Section 5.3, which leads to the regression envelopes developed in Section 5.4. Section 5.5 includes various envelope regression applications: weighted least squares, generalized linear models and Cox regression. Simulation results are given in Section 5.6, and illustrative analysis are given in Section 5.8. Although our focus in this article is on vector-valued parameters, we describe in Section 5.2.4 how envelopes for matrix parameters can be constructed generally. Proofs and other technical details are included in Section 5.9.

5.2 A general definition of envelopes

Envelopes arose in the studies reviewed in Section 5.1 as natural consequences of postulating the presence of immaterial information in $\mathbf{Y}$ or $\mathbf{X}$. However, they provide no guidance on how to employ parallel reasoning in more complex settings, like generalized linear models, or in settings without a clear regression structure. In terms of Definition 1.2, the previous studies offer no guidance on how to choose the matrix $\mathbf{M}$ and the subspace $\mathcal{B}$ for use in general multivariate problems, particularly since there are many ways to represent the same envelope, as indicated in parts 2 and 3 of Proposition 1.1. In this section, we propose a broad criterion to guide these selections.

5.2.1 Enveloping a vector-valued parameter

Let $\widehat{\boldsymbol{\theta}}$ denote an estimator of a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ based on a sample of size n. Let $\boldsymbol{\theta}_t$ denote the true value of $\boldsymbol{\theta}$ and assume, as is often the case, that $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_t)$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\mathbf{V}(\boldsymbol{\theta}_t) > 0$ as $n \to \infty$. To accommodate the presence of nuisance parameters we decompose $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\psi}^T, \boldsymbol{\phi}^T)^T$, where $\boldsymbol{\phi} \in \mathbb{R}^q$, $q \leq m$, is the parameter vector of interest and $\boldsymbol{\psi} \in \mathbb{R}^{m-q}$ is the nuisance parameter vector. The asymptotic covariance matrix of $\widehat{\boldsymbol{\phi}}$ is represented as $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, which is the $q \times q$ lower right block of $\mathbf{V}(\boldsymbol{\theta}_t)$. Then we define an envelope as follows.

Definition 5.1. *The envelope for the parameter $\boldsymbol{\phi} \in \mathbb{R}^q$ is defined as $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) \subseteq \mathbb{R}^q$.*

This definition of an envelope expands previous approaches reviewed in Section 5.1 in a variety of ways. First, it links the envelope to a particular pre-specified method of estimation through the covariance matrix $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, while in previous approaches the method of estimation played only a background role. The goal of an envelope is to improve that pre-specified estimator, perhaps a maximum likelihood, least squares or robust estimator, depending on the original goals of the analysis. Second, the matrix to be reduced – here $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ – is dictated by the method of estimation. Third, the matrix to be reduced can now depend on the parameter being estimated, in addition to perhaps other parameters.

Definition 5.1 reproduces the partial envelopes for $\boldsymbol{\beta}_1$ in Cook and Su (2011) and the envelopes for $\boldsymbol{\beta}$ when it is a vector; that is, when $r = 1$ and $p > 1$ or when $r > 1$ and $p = 1$. It also reproduces the the partially maximized log likelihood function (5.1.3) by setting $\mathbf{M} = \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ and $\mathbf{U} = \boldsymbol{\phi}_t\boldsymbol{\phi}_t^T$. To apply Definition 5.1 for the partial envelope of $\boldsymbol{\beta}_1$ based on model $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\epsilon}$, the asymptotic covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}_1$ is $\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, where $(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}$ is the $(1,1)$ element of the inverse of $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T$. Consequently, by Proposition 1.1, $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}}(\boldsymbol{\beta}_1) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, and thus Definition 5.1 recovers the partial envelopes of Su and Cook (2011). To construct the partially maximized log likelihood (5.1.3) we set $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}$ and $\mathbf{U} = \boldsymbol{\beta}_1\boldsymbol{\beta}_1^T$. Then using sample versions gives

$$\begin{aligned} J_n(\boldsymbol{\Gamma}) &= \log|(\mathbf{S}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Gamma}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}^T| + \log|\boldsymbol{\Gamma}\{(\mathbf{S}_{\mathbf{X}}^{-1})_{11}\mathbf{S}_{\mathbf{Y}|\mathbf{X}} + \widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_1^T\}^{-1}\boldsymbol{\Gamma}^T| \\ &= \log|\boldsymbol{\Gamma}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}\boldsymbol{\Gamma}^T| + \log|\boldsymbol{\Gamma}\mathbf{S}_{\mathbf{Y}|\mathbf{X}_2}^{-1}\boldsymbol{\Gamma}^T|, \end{aligned}$$

which is the partially maximized log-likelihood of Su and Cook (2011). It is important to note that, although $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}}(\boldsymbol{\beta}_1) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}_1)$, Definition 5.1 requires that we use $\mathbf{V}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} = (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})_{11}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and not $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ alone.

As a second illustration, consider $\mathbf{X}$ reduction in model (5.1.1) with $r = 1$ and $p > 1$. To emphasize the scalar response, let $\sigma^2_{Y|\mathbf{X}} = \text{var}(\varepsilon)$ with sample residual variance $s^2_{Y|\mathbf{X}}$. The ordinary least squares estimator of $\boldsymbol{\beta}$ has asymptotic variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2_{Y|\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$. Direct application of Definition 5.1 then leads to the $\sigma^2_{Y|\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$-envelope of span$(\boldsymbol{\beta})$, $\mathcal{E}_{\sigma^2_{Y|\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}}(\boldsymbol{\beta})$. However, it follows from Proposition 1.1 that this envelope is equal to $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, which is the envelope used by Cook, et al., (2013) when establishing connections with partial least squares. To construct the corresponding version of (5.1.3), let $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $\mathbf{U} = \boldsymbol{\beta}\boldsymbol{\beta}^T$. Then substituting sample quantities

$$\begin{aligned} J_n(\boldsymbol{\Gamma}) &= \log|s^2_{Y|\mathbf{X}}\boldsymbol{\Gamma}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}^T| + \log|\boldsymbol{\Gamma}\{s^2_{Y|\mathbf{X}}\mathbf{S}_{\mathbf{X}}^{-1} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T\}^{-1}\boldsymbol{\Gamma}^T| \\ &= \log|\boldsymbol{\Gamma}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}^T| + \log|\boldsymbol{\Gamma}(\mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{X}Y}\mathbf{S}_{\mathbf{X}Y}^T/s^2_Y)^{-1}\boldsymbol{\Gamma}^T|, \end{aligned} \tag{5.2.1}$$

which gives the partially maximized log-likelihood of Cook et al. (2013). Although $\mathcal{E}_{\sigma^2_{Y|\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, it is again important that we use $\mathbf{M} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2_{Y|\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ in the construction of $J_n(\boldsymbol{\Gamma})$.

Definition 5.1 in combination with $J(\boldsymbol{\Gamma})$ can also be used to derive envelope estimators for new problems. For example, consider enveloping for a multivariate mean $\boldsymbol{\mu}$ in the model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}_{\mathbf{Y}})$. We take $\boldsymbol{\phi} = \boldsymbol{\mu}$ and $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \mathbf{Y}_i$. Then $\mathbf{M} = \mathbf{V}_{\boldsymbol{\mu}\boldsymbol{\mu}} = \boldsymbol{\Sigma}_{\mathbf{Y}}$, which is the asymptotic covariance matrix of $\widehat{\boldsymbol{\mu}}$, $\mathbf{U} = \boldsymbol{\mu}\boldsymbol{\mu}^T$, and $\mathbf{M} + \mathbf{U} = E(\mathbf{Y}\mathbf{Y}^T)$. Substituting sample versions of $\mathbf{M}$ and $\mathbf{U}$ leads to the same objective function $J_n(\boldsymbol{\Gamma})$ as that obtained when deriving the envelope estimator from scratch.

5.2.2 Estimation in general

Having seen that Definition 5.1 recovers past envelopes, we next turn to its general use in estimation. The function $J_n(\mathbf{\Gamma})$ in (5.1.3) can be used as a generic moment-based objective function requiring only matrices $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$. Consequently, distributional assumptions, such as normality, are not strict requirement for estimators based on $J_n(\mathbf{\Gamma})$ to be useful, a conclusion that is supported by previous work and by our experience. Nevertheless, optimization of $J_n(\mathbf{\Gamma})$ becomes computationally difficult as the dimensions involved increase. For instance, consider the envelope estimator of $\boldsymbol{\beta}$ based on estimating a basis for $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ with dimension u. Since $J_n(\mathbf{\Gamma}) = J_n(\mathbf{\Gamma O})$ for any orthogonal marix $\mathbf{O} \in \mathbb{R}^{u\times u}$, optimization of J is over the Grassmannian $\mathcal{G}_{u,r}$. Since $u(r-u)$ real numbers are required to specify an element of $\mathcal{G}_{u,r}$ uniquely, optimization of J is essentially over $u(r-u)$ real dimensions, and can be time consuming and sensitive to starting values when this dimension is large. The 1D algorithm in Chapter 2 mitigates these computational issues. Setting $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\phi}}\widehat{\boldsymbol{\phi}}^T \in \mathbb{S}^{q\times q}$ and $\widehat{\mathbf{M}}$ equal to a $\sqrt{n}$-consistent estimator of $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) \in \mathbb{S}^{q\times q}$, the 1D algorithm will estimate a basis of the envelope by $\widehat{\mathbf{G}}_u$. it follows from Proposition 2.5 that the 1D algorithm provides a $\sqrt{n}$-consistent estimator $\widehat{\mathbf{P}}_u = \widehat{\mathbf{G}}_u\widehat{\mathbf{G}}_u^T$ of the projection onto $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) \subseteq \mathbb{R}^p$, assuming that $u = \dim(\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t))$ is known. The *moment-based envelope estimator* $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\mathbf{P}}_u\widehat{\boldsymbol{\phi}}$ is then a $\sqrt{n}$-consistent estimator of $\boldsymbol{\phi}_t$.

To gain intuition about the potential gains of the envelope estimator, assume that a basis $\mathbf{\Gamma}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is known and write the envelope estimator as $\mathbf{P}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}$. Then since the envelope reduces $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, we have

$$\begin{aligned}
\text{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}) &= \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) = \mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\mathbf{\Gamma}}, \\
\text{avar}(\sqrt{n}\mathbf{P}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}) &= \mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\mathbf{\Gamma}} \le \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t).
\end{aligned}$$

These relationships allow some straightforward intuition by writing $\sqrt{n}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_t) = \sqrt{n}(\mathbf{P}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_t) + \sqrt{n}\mathbf{Q}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}$. The second term $\sqrt{n}\mathbf{Q}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}$ is asymptotically normal with mean 0 and variance $\mathbf{Q}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\mathbf{\Gamma}}$, and is asymptotically independent of $\sqrt{n}(\mathbf{P}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_t)$. Consequently, we think of $\mathbf{Q}_{\mathbf{\Gamma}}\widehat{\boldsymbol{\phi}}$ as the immaterial information in $\widehat{\boldsymbol{\phi}}$. The envelope estimator then achieves efficiency gains by essentially eliminating the immaterial variation $\mathbf{Q}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\mathbf{\Gamma}}$, the greatest gains being achieved when $\mathbf{Q}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{Q}_{\mathbf{\Gamma}}$ is large relative to $\mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\mathbf{\Gamma}}$. Of course, we will typically need to estimate $\mathbf{\Gamma}$ in practice, which will mitigate the asymptotic advantages when $\mathbf{\Gamma}$ is known. But when the immaterial variation is larger compare to the "cost" of estimating the envelope, substantial gain will still be achieved.

If $\widehat{\boldsymbol{\phi}}$ is obtained by minimizing a model-based objective function $\widehat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathbb{R}^q} F_n(\boldsymbol{\phi})$ then, as an alternative to the moment based estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\mathbf{P}}_u\widehat{\boldsymbol{\phi}}$, we prefer to construct the envelope estimator as $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\mathbf{G}}_u\widehat{\boldsymbol{\eta}}$, where $\widehat{\mathbf{G}}_u$ is obtained from the sample version of the 1D algorithm and $\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^u} F_n(\widehat{\mathbf{G}}_u\boldsymbol{\eta})$. Sometimes these two approaches are identical, i.e., response envelopes (Cook et al. 2010) and partial envelopes (Su and Cook

2011). We elaborate on the objective function approach in Section 5.3 in the context of maximum likelihood estimators.

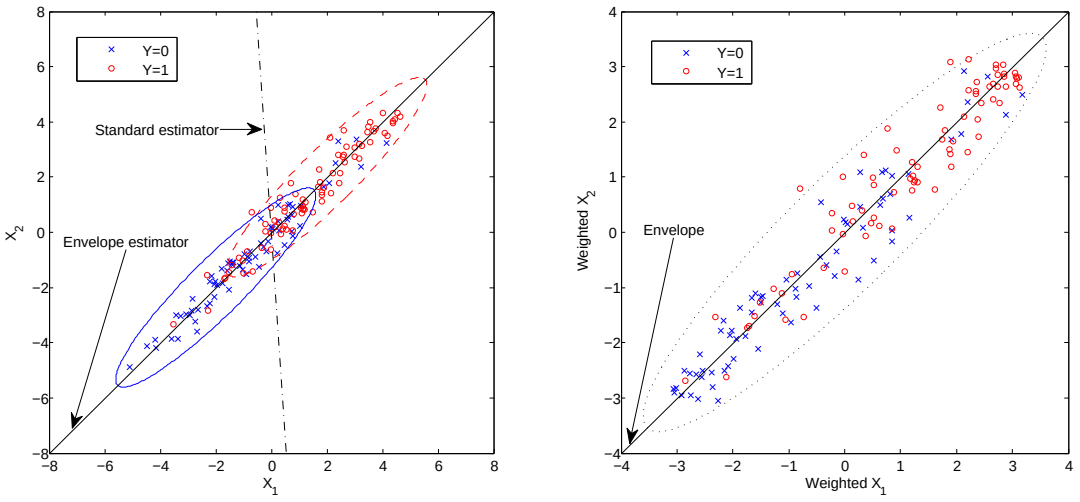
Although we do not have an expression for the variance of the moment-based estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\mathbf{P}}_u\widehat{\boldsymbol{\phi}}$, our simulation results indicate that the bootstrap is a reliable method for estimating it. Depending on context, cross validation, an information criteria like AIC and BIC, or sequential hypothesis testing can be used to aid selection of u, as Cook et al. (2010, 2013) and Su and Cook (2011).

5.2.3 Envelope in logistic regression

Following Definition 5.1, we derive in Section 5.5.2 that the envelope in logistic regression is $\mathcal{E}_{\boldsymbol{\Sigma}_{(W)\mathbf{X}}}(\boldsymbol{\beta})$, where $\boldsymbol{\Sigma}_{(W)\mathbf{X}} = \text{cov}(\sqrt{W}\mathbf{X})$ is the covariance of the weighted predictors $\sqrt{W}\mathbf{X}$ and $W = \exp(\boldsymbol{\beta}^T\mathbf{X})/\{1+\exp(\boldsymbol{\beta}^T\mathbf{X})\}^2$. This is the first time that the idea of envelopes is extended beyond linear regression. As introduction and motivation, we now use a small simulation study to illustrate the advantages of envelopes.

We generated 150 independent observations as follows: $Y_i|\mathbf{X}_i \sim \text{Bernoulli}(\text{logit}(\boldsymbol{\beta}^{\text{T}}\mathbf{X}_{\text{i}}))$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T = (0.25, 0.25)^T$ and $\mathbf{X}_i \sim N(0, \boldsymbol{\Sigma}_{\mathbf{X}})$. We let $\text{span}(\boldsymbol{\beta})$ be the principal eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}$ with eigenvalue 10 and let the other eigenvalue be 0.1, which led to co-linearity and made the estimation challenging. We plotted the data in Figure 5.1(a), in which we used two lines to indicate the directions of the two estimators: standard estimator $\widehat{\boldsymbol{\beta}} = (-0.047, 0.692)^T$ with standard errors $\text{SE}(\widehat{\boldsymbol{\beta}}) = (0.418, 0.429)^T$, and envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = (0.323, 0.317)^T$ with standard errors $\text{SE}(\widehat{\boldsymbol{\beta}}_{\text{env}}) = (0.058, 0.057)^T$. The components of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ are large relative to their standard errors, while the components of $\widehat{\boldsymbol{\beta}}$ are not. The envelope estimator was obtained by maximizing the likelihood over $\eta \in \mathbb{R}^1$ according to the reparameterization $\boldsymbol{\beta} = \widehat{\mathbf{G}}_u\eta$, where $\widehat{\mathbf{G}}_u \in \mathbb{R}^2$ was the estimated envelope basis from the 1D algorithm and $u = 1$. The moment-based envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env},2} = \widehat{\mathbf{P}}_u\widehat{\boldsymbol{\beta}} = (0.322, 0.316)^T$ was very close to the likelihood-based estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$. We also plotted the weighted predictors $\sqrt{W}\mathbf{X}$ in Figure 5.1 (b) so that we can see from the figure that the envelope is the principal eigenvector of $\boldsymbol{\Sigma}_{(W)\mathbf{X}}$.

To further demonstrate the point that standard logistic regression estimator are highly variable in the presence of co-linearity, we simulated another data set according to the same model. Again, the envelope estimator stayed close to the truth, $\widehat{\boldsymbol{\beta}}_{\text{env}} = (0.208, 0.209)^T$ with standard errors $\text{SE}(\widehat{\boldsymbol{\beta}}_{\text{env}}) = (0.047, 0.047)^T$. And again the moment-based envelope estimator was very similar, $\widehat{\boldsymbol{\beta}}_{\text{env},2} = (0.207, 0.208)^T$. But the standard estimator varied substantially, $\widehat{\boldsymbol{\beta}} = (0.665, -0.248)^T$ with standard errors $\text{SE}(\widehat{\boldsymbol{\beta}}) = (0.401, 0.399)^T$.



(a) Estimated envelope and standard estimators.

(b) Visualization of $\mathbf{\Sigma}_{(W)\mathbf{X}} = \mathrm{cov}(\sqrt{W}\mathbf{X})$.

Figure 5.1: Illustration of envelopes in logistic regression: (a) envelope and standard estimators in a simulated data set, the true envelope $\mathcal{E}_{\mathbf{\Sigma}_{(W)\mathbf{X}}}(\boldsymbol{\beta}_t)$ and its sample estimate are indistinguishable in the plot; (b) visualization of the covariance $\mathbf{\Sigma}_{(W)\mathbf{X}} = \mathrm{cov}(\sqrt{W}\mathbf{X})$ by plotting the weighted predictors $\sqrt{W}\mathbf{X}$ on axes, where the true weights were used: $W_i = \exp(\boldsymbol{\beta}_t^T\mathbf{X}_i)/(1+\exp(\boldsymbol{\beta}_t^T\mathbf{X}_i))^2$.

5.2.4 Enveloping a matrix-valued parameter

Definition 5.1 is sufficiently general to cover a matrix-valued parameter $\boldsymbol{\phi} \in \mathbb{R}^{r\times c}$ by considering the avar$(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\phi}}))$-envelope of $\mathrm{vec}(\boldsymbol{\phi})$. However, matrix-valued parameters come with additional structure that is often desirable to maintain during estimation. For instance, in the multivariate linear model reviewed in Section 5.1, envelope construction was constrained to reflect separate row and column reduction of the matrix parameter $\boldsymbol{\beta}$. The advantages of row and column reductions have been discussed by Li et al. (2010) and Hung et al. (2012). Although our primary focus is on vector-valued parameters, in this section we indicate how to adapt for a matrix-valued parameter.

Suppose that $\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_t)$ converges to a matrix normal distribution with mean 0, column variance $\mathbf{\Delta}_L \in \mathbb{S}^{r\times r}$ and row variance $\mathbf{\Delta}_R \in \mathbb{S}^{c\times c}$. (See Dawid (1981) for background on the matrix normal distribution.) Then

$$\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{\phi}}) - \mathrm{vec}(\boldsymbol{\phi}_t)) \to N(0, \mathbf{\Delta}_R \otimes \mathbf{\Delta}_L), \tag{5.2.2}$$

and direct application of Definition 5.1 yields the envelope $\mathcal{E}_{\mathbf{\Delta}_R\otimes\mathbf{\Delta}_L}(\mathrm{vec}(\boldsymbol{\phi}_t))$. However, this envelope may not preserve the intrinsic row-column structure of $\boldsymbol{\phi}_t$. In the following definition we introduce a restricted class of envelopes that maintain the matrix structure of $\boldsymbol{\phi}_t$.

Definition 5.2. *Assume that $\widehat{\boldsymbol{\phi}}$ is asymptotically matrix normal as give in (5.2.2). Then the tensor envelope for $\boldsymbol{\phi}_t$, denoted by $\mathcal{K}_{\mathbf{\Delta}_R\otimes\mathbf{\Delta}_L}(\boldsymbol{\phi}_t)$, is defined as the intersection of*

all reducing subspaces $\mathcal{E}$ *of* $\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L$ *that contain* $\text{span}(\text{vec}(\boldsymbol{\phi}_t))$ *and can be written as* $\mathcal{E} = \mathcal{E}_R \otimes \mathcal{E}_L$ *with* $\mathcal{E}_R \subseteq \mathbb{R}^c$ *and* $\mathcal{E}_L \subseteq \mathbb{R}^r$.

We see from this definition that $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t)$ always exist and is the smallest subspace with the required properties. Let $\mathbf{L} \in \mathbb{R}^{r \times d_c}$, $d_c < r$, and $\mathbf{R} \in \mathbb{R}^{c \times d_r}$, $d_r < c$ be semi-orthogonal matrices such that $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) = \text{span}(\mathbf{R} \otimes \mathbf{L})$. By definition, $\text{span}(\boldsymbol{\phi}_t) \subseteq \text{span}(\mathbf{L})$ and $\text{span}(\boldsymbol{\phi}_t^T) \subseteq \text{span}(\mathbf{R})$. Hence we have $\boldsymbol{\phi}_t = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$ and $\text{vec}(\boldsymbol{\phi}_t) = (\mathbf{R} \otimes \mathbf{L})\text{vec}(\boldsymbol{\eta})$ for some $\boldsymbol{\eta} \in \mathbb{R}^{d_c \times d_r}$.

The next proposition shows how to factor $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) \in \mathbb{R}^{rc}$ into the tensor product of envelopes $\mathcal{E}_{\boldsymbol{\Delta}_R}(\boldsymbol{\phi}_t^T) \in \mathbb{R}^r$ and $\mathcal{E}_{\boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) \in \mathbb{R}^c$ for the row and column spaces of $\boldsymbol{\phi}_t$. These tensor factors are envelopes in smaller spaces that preserve the row and column structure and can facilitate analysis and interpretation.

Proposition 5.1. $\mathcal{K}_{\boldsymbol{\Delta}_R \otimes \boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t) = \mathcal{E}_{\boldsymbol{\Delta}_R}(\boldsymbol{\phi}_t^T) \otimes \mathcal{E}_{\boldsymbol{\Delta}_L}(\boldsymbol{\phi}_t)$.

For example, in reference to model (5.1.1), the distribution of $\widehat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{XY}}^T \mathbf{S}_{\mathbf{X}}^{-1}$ satisfies (5.2.2) with $\boldsymbol{\Delta}_R = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Delta}_L = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The tensor envelope is then

$$\mathcal{K}_{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}}(\boldsymbol{\beta}^T) \otimes \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta}).$$

If we are interested in reducing only the column space of $\boldsymbol{\beta}$, which corresponds to response reduction, we would use $\mathbb{R}^p \otimes \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$ for constructing an envelope estimator of $\boldsymbol{\beta}$, and then $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\eta}\mathbf{I}_p = \mathbf{L}\boldsymbol{\eta}$ where $\mathbf{L}$ is a semi-orthogonal basis for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\boldsymbol{\beta})$, which reproduces the envelope construction in Cook et al. (2010). Similarly, if we are interested in only predictor reduction, we would take $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T) \otimes \mathbb{R}^r$, which reproduces the envelope construction in Cook et al. (2013). More generally, the definition of the tensor envelope and Proposition 5.1 straightforwardly connects and combines the envelope models in the predictor space and in the response space, leading to the simultaneous envelope method in Cook and Zhang (2014).

5.3 Envelopes for maximum likelihood estimators

While the setup of Section 5.2 was quite broad, different and more suitable envelope methods may emerge in narrower contexts. In this section we narrow the context of our study to likelihood-based estimators.

Consider estimating $\boldsymbol{\theta}$ as $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L_n(\mathbf{D}, \boldsymbol{\theta})$, where $L_n(\mathbf{D}, \boldsymbol{\theta})$ is a log-likelihood that is twice continuously differentiable in an open neighborhood of $\boldsymbol{\theta}_t$ and $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_n)^T$ is a generic representation of the data. We will often suppress the data and write $L_n(\mathbf{D}, \boldsymbol{\theta})$ more compactly as $L_n(\boldsymbol{\theta})$. Then under standard conditions $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)$ is asymptotically normal with mean 0 and covariance matrix $\mathbf{V}(\boldsymbol{\theta}_t) = \mathbf{F}^{-1}(\boldsymbol{\theta}_t)$, where $\mathbf{F}$ is the Fisher information matrix for $\boldsymbol{\theta}$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\phi}}$, $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$, is

the lower right block of $\mathbf{V}(\boldsymbol{\theta}_t)$. Let $\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ be the $p \times p$ lower right block of $\mathbf{F}(\boldsymbol{\theta}_t)$. It follows that $\mathcal{E}_{\mathbf{V}(\boldsymbol{\theta}_t)}(\boldsymbol{\theta}_t) = \mathcal{E}_{\mathbf{F}(\boldsymbol{\theta}_t)}(\boldsymbol{\theta}_t)$. However, $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) \neq \mathcal{E}_{\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ in general but $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t) = \mathcal{E}_{\mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ when $\mathbf{F}(\boldsymbol{\theta}_t)$ is block-diagonal, $\mathbf{F}(\boldsymbol{\theta}_t) = \mathrm{diag}(\mathbf{F}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_t), \mathbf{F}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t))$.

To see the potential advantages of envelopes in the context of maximum likelihood estimators, assume that we know an orthogonal basis $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$, where $u = \dim(\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t))$. Since $\boldsymbol{\phi}_t \in \mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$, we have $\boldsymbol{\phi}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_t$ for $\boldsymbol{\eta}_t \in \mathbb{R}^u$. Consequently, for fixed $\boldsymbol{\Gamma}$, we can write the log-likelihood as $L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\psi}, \boldsymbol{\phi}) = L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta})$, and the envelope estimators become

$$
\begin{aligned}
(\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}) &= \arg\max_{\boldsymbol{\psi},\boldsymbol{\eta}} L_n(\boldsymbol{\psi}, \boldsymbol{\Gamma}\boldsymbol{\eta}), && (5.3.1)\\
\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} && (5.3.2)\\
\widehat{\boldsymbol{\theta}}_{\boldsymbol{\Gamma}} &= (\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}^T, \widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}^T)^T, && (5.3.3)
\end{aligned}
$$

Since $\boldsymbol{\phi} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, any basis for $\mathcal{E}_{\mathbf{V}}(\boldsymbol{\phi}_t)$ will give the same solution $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$: for an orthogonal matrix $\mathbf{O}$, write $\boldsymbol{\phi} = \boldsymbol{\Gamma}\mathbf{O}\mathbf{O}^T\boldsymbol{\eta}$. Then $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}\mathbf{O}} = \mathbf{O}^T\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$.

The estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$ given in (5.3.2) is in general different from the estimator $\mathbf{P}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\phi}}$ discussed near the end of Section 5.2.2. However, as implied by the following proposition, these two estimators have the same asymptotic distribution, which provides some support for the simple projection estimator of Section 5.2.2.

Proposition 5.2. *As $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} - \boldsymbol{\phi}_t)$ converges to a normal random vector with mean 0 and asymptotic covariance*

$$
\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\mathbf{P}_{\boldsymbol{\Gamma}} \leq \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t).
$$

The $\sqrt{n}$-consistent nuisance parameter estimator also satisfies $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}) \leq \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}})$.

The following corollary to Proposition 5.2 characterizes the asymptotic variance when an arbitrary envelope is used.

Corollary 5.1. *If $\boldsymbol{\Gamma}$ is a basis for an arbitrary envelope $\mathcal{E}_{\mathbf{M}}(\boldsymbol{\phi}_t)$, where $\mathbf{M}$ is a symmetric positive definite matrix, then*

$$
\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) = \boldsymbol{\Gamma}\{\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}\}^{-1}\boldsymbol{\Gamma}^T \leq \mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t). \tag{5.3.4}
$$

The above expression shows that $\boldsymbol{\Gamma}$ is intertwined with $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)$ in the asymptotic covariance of the envelope estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$, while the envelope by Definition 5.1 makes $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}})$ more interpretable because the material and the immaterial variations are separable.

As formulated, this likelihood context allows us to construct the envelope estimator $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}$ when a basis $\boldsymbol{\Gamma}$ for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is known, but it does not by itself provide a basis estimator $\widehat{\boldsymbol{\Gamma}}$. However, a basis can be estimated by using the 1D algorithm (Algorithm 1),

setting $\mathbf{M} = \mathbf{V}_{\phi\phi}(\boldsymbol{\theta})$ and $\mathbf{U} = \boldsymbol{\phi}\boldsymbol{\phi}^T$ and plugging-in the pre-specified estimator $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta})$ to get $\sqrt{n}$-consistent estimators of $\mathbf{M}$ and $\mathbf{U}$. The envelope estimator is then $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\boldsymbol{\phi}}_{\widehat{\boldsymbol{\Gamma}}}$, where $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{G}}_u$. Then we have the following proposition.

Proposition 5.3. *If the estimated basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ is obtained by the 1D algorithm (Algorithm 1), then the envelope estimator $\widehat{\boldsymbol{\phi}}_{\text{env}} = \widehat{\boldsymbol{\phi}}_{\widehat{\boldsymbol{\Gamma}}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\phi}$.*

The envelope estimator $\widehat{\boldsymbol{\phi}}_{\text{env}}$ depends on the pre-specified estimator $\widehat{\boldsymbol{\theta}}$ through $\mathbf{M} = \mathbf{V}_{\phi\phi}(\boldsymbol{\theta})$ and $\mathbf{U} = \boldsymbol{\phi}\boldsymbol{\phi}^T$. Although it is $\sqrt{n}$-consistent, we have found empirically that, when $\mathbf{M}$ depends non-trivially on $\boldsymbol{\theta}_t$, it can often be improved by iterating so the current estimate of $\widehat{\boldsymbol{\theta}}$ is use to construct estimates of $\mathbf{M}$ and $\mathbf{U}$. The iteration can be implemented as follows. Initialize $\widehat{\boldsymbol{\theta}}_0 = \widehat{\boldsymbol{\theta}}$, and let $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{U}}_k$ be the estimators of $\mathbf{M}$ and $\mathbf{U}$ based on the k-th envelope estimator $\widehat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}$, so $\widehat{\mathbf{M}}_0$ and $\widehat{\mathbf{U}}_0$ are based only on the pre-specified estimator. Then for $k = 0, 1, \ldots,$

1. Using the 1D algorithm, construct an estimated basis $\widehat{\boldsymbol{\Gamma}}_k$ for $\mathcal{E}_{\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)}(\boldsymbol{\phi}_t)$ using $\widehat{\mathbf{M}}_k = \mathbf{V}_{\phi\phi}(\widehat{\boldsymbol{\theta}}_k)$ and $\widehat{\mathbf{U}}_k = \widehat{\boldsymbol{\phi}}_k\widehat{\boldsymbol{\phi}}_k^T$.
2. Set $\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_{\widehat{\boldsymbol{\Gamma}}_k}$ from (5.3.3).
3. Stop if a measure of the change between $\widehat{\boldsymbol{\phi}}_{k+1}$ and $\widehat{\boldsymbol{\phi}}_k$ is sufficiently small; otherwise, return to the first step.

5.4 Regression

In this section we further narrow our study by adding a regression structure to the likelihood. This allows us to use likelihood-based estimators of the envelope instead of the 1D algorithm, which provides a further refinement of envelope methodology.

5.4.1 Conditional and unconditional inference in regression

Let $Y \in \mathbb{R}^1$ and $\mathbf{X} \in \mathbb{R}^p$ have a joint distribution with parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\psi}^T)^T \in \mathbb{R}^{q+p+s}$, so the joint density or mass function can be written as $f(Y, \mathbf{X}|\boldsymbol{\theta}) = g(Y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X})h(\mathbf{X}|\boldsymbol{\psi})$ and the observed data are $\mathbf{D}_i \equiv (Y_i, \mathbf{X}_i^T)^T \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$. We take $\boldsymbol{\beta}$ to be the parameter vector of interest and, prior to the introduction of envelopes, we restrict the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ to a product space. The predictors $\mathbf{X}$ are ancillary in most regressions and thus analysis is often based on the conditional likelihood. Let $L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(Y_i, \mathbf{X}_i|\boldsymbol{\theta})$ be the full log-likelihood, let the conditional log-likelihood be represented by $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^n \log g(Y_i|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X}_i)$ and let $M_n(\boldsymbol{\psi}) = \sum_{i=1}^n \log h(\mathbf{X}_i|\boldsymbol{\psi})$ be the marginal log-likelihood for $\boldsymbol{\psi}$. Then we can decompose $L_n(\boldsymbol{\theta}) = C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) + M_n(\boldsymbol{\psi})$. Since our primary interest lies in $\boldsymbol{\beta}$ and $\mathbf{X}$ is ancillary, estimators are typically obtained

as

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} C_n(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{5.4.1}$$

Our goal here is to improve the pre-specified estimator $\widehat{\boldsymbol{\beta}}$ by introducing the envelope $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$, where $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t) = \text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}})$.

Let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{p\times p}$ denote an orthogonal basis for $\mathbb{R}^p$ where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ is a basis for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$. Since $\boldsymbol{\beta}_t \in \mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$, we can write $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_t$ for some $\boldsymbol{\eta}_t \in \mathbb{R}^u$. Because $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t)$ typically depends on the distribution of $\mathbf{X}$ and $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ reduces $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_t)$, the marginal M_n will depend on $\boldsymbol{\Gamma}$. Then the log-likelihood becomes $L_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\psi}_1, \boldsymbol{\Gamma}) = C_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) + M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})$, where $\boldsymbol{\psi}_1$ represents any parameters remaining after incorporating $\boldsymbol{\Gamma}$. Since both C_n and M_n depend on $\boldsymbol{\Gamma}$, the predictors are no longer ancillary after incorporated the envelope structure and estimation must be carried out by maximizing $\{C_n(\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) + M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})\}$.

The relationship between $\mathbf{X}$ and $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ that is embodied in $M_n(\boldsymbol{\psi}_1, \boldsymbol{\Gamma})$ could be complicated, depending on the distribution of $\mathbf{X}$. However, as described in the following proposition, it simplifies considerably when $\text{E}(\mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{X})$ is a linear function of $\boldsymbol{\Gamma}^T\mathbf{X}$. This is well-known as the linearity condition in the sufficient dimension reduction literature where there $\boldsymbol{\Gamma}$ denotes a basis for the central subspace (Cook 1996). Background on the linearity condition, which is widely regarded as restrictive but nonetheless rather mild, is available from Cook (1998), Li and Wang (2007) and many other articles in sufficient dimension reduction. For instance, if $\mathbf{X}$ follows an elliptically contoured distribution, the linearity condition will be guaranteed for any $\boldsymbol{\Gamma}$ (Eaton 1986).

Proposition 5.4. *Assume that* $\text{E}(\mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{X})$ *is a linear function of* $\boldsymbol{\Gamma}^T\mathbf{X}$. *Then* $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_t)$.

One implication of Proposition 5.4 is that $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_t)$ can be used as a replacement for $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ when developing likelihood-based envelop methods, although as mentioned previously this replacement is not appropriate when using the 1D algorithm as a general method of construction. A second implication is that, for some positive definite matrices $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$, $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ and thus M_n must depend on $\boldsymbol{\Gamma}$ through the marginal covariance $\boldsymbol{\Sigma}_\mathbf{X}$. Consequently, we can write $M_n(\boldsymbol{\Sigma}_\mathbf{X}, \boldsymbol{\psi}_2) = M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\psi}_2)$, where $\boldsymbol{\psi}_2$ represents any remaining parameters in the marginal function. If $\mathbf{X}$ is normal with mean $\boldsymbol{\mu}_\mathbf{X}$ and variance $\boldsymbol{\Sigma}_\mathbf{X}$, then $\boldsymbol{\psi}_2 = \boldsymbol{\mu}_\mathbf{X}$ and $M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\psi}_2) = M_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\mu}_\mathbf{X})$ is the marginal normal log-likelihood. In this case, it is possible to maximize M_n over all its parameters except $\boldsymbol{\Gamma}$, as stated in the following lemma.

Lemma 5.1. *Assume that* $\mathbf{X} \in \mathbb{R}^p$ *is multivariate normal* $N(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ *and that* $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ *is a semi-orthogonal basis matrix for* $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_t)$. *Then* $\widehat{\boldsymbol{\mu}}_{\mathbf{X},\boldsymbol{\Gamma}} = \overline{\mathbf{X}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{S}_\mathbf{X}\mathbf{P}_{\boldsymbol{\Gamma}} +$

$\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\boldsymbol{\Gamma}}$ *and*

$$\begin{aligned} M_n(\boldsymbol{\Gamma}) &:= \max_{\boldsymbol{\Omega},\boldsymbol{\Omega}_0,\boldsymbol{\mu}_{\mathbf{X}}} M_n(\boldsymbol{\Gamma},\boldsymbol{\Omega},\boldsymbol{\Omega}_0,\boldsymbol{\mu}_{\mathbf{X}}) \qquad (5.4.2) \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}_0|\right\} \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}, \qquad (5.4.3) \end{aligned}$$

where $(\boldsymbol{\Gamma},\boldsymbol{\Gamma}_0) \in \mathbb{R}^{p\times p}$ *is an orthogonal basis for* $\mathbb{R}^p$*. Moreover, the global maximum of* $M_n(\boldsymbol{\Gamma})$ *is attained at all subsets of* u *eigenvectors of* $\mathbf{S}_{\mathbf{X}}$*.*

This lemma indicates that if $\mathbf{X}$ is marginally normal, the envelope estimators are then constructed from

$$(\widehat{\boldsymbol{\alpha}}_{\text{env}},\widehat{\boldsymbol{\eta}},\widehat{\boldsymbol{\Gamma}}) = \arg\max\{C_n(\boldsymbol{\alpha},\boldsymbol{\eta},\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma})\}. \qquad (5.4.4)$$

In particular, the envelope estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$ and, from Lemma 5.1, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\text{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$. It follows also from Lemma 5.1 that one role of M_n is to pull $\widehat{\boldsymbol{\Gamma}}$ toward the reducing subspaces of $\mathbf{S}_{\mathbf{X}}$, although it will not necessarily coincide with any such subspace.

5.4.2 Asymptotic properties with normal predictors

In this section we describe the asymptotic properties of envelope estimators when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal parameter vectors and $\mathbf{X}$ is normally distributed. Cox and Reid (1987) discuss the construction and interpretation of orthogonal parameters, and we describe the construction of orthogonal parameters in the context of generalized linear models in Section 5.5.2. We also contrast the asymptotic behavior of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}}$ with that of the estimator $\widehat{\boldsymbol{\beta}}$ from $C_n(\boldsymbol{\alpha},\boldsymbol{\beta})$, and comment on other settings at the end of the section.

The parameters involved in the coordinate representation of the envelope model are $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Gamma}$. Since the parameters $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ depend on the basis $\boldsymbol{\Gamma}$ and the objective function is invariant under orthogonal transformations of $\boldsymbol{\Gamma}$, the estimators of these parameters are not unique. Hence, we consider only the asymptotic properties of the estimable functions $\boldsymbol{\alpha}$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, which are invariant under choice of basis $\boldsymbol{\Gamma}$ and thus have unique maximizers. Under the normality assumption, $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\Sigma}_{\mathbf{X}})$, we neglect the mean vector $\boldsymbol{\mu}_{\mathbf{X}}$ since it is orthogonal to all of the other parameters. We define the following parameters $\boldsymbol{\phi}$ and estimable functions h.

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \\ \boldsymbol{\phi}_5 \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\eta} \\ \text{vec}(\boldsymbol{\Gamma}) \\ \text{vech}(\boldsymbol{\Omega}) \\ \text{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \quad \boldsymbol{h} = \begin{pmatrix} \boldsymbol{h}_1(\boldsymbol{\phi}) \\ \boldsymbol{h}_2(\boldsymbol{\phi}) \\ \boldsymbol{h}_3(\boldsymbol{\phi}) \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}) \end{pmatrix}.$$

Since the number of free parameters in $\boldsymbol{h}$ is $q + p + p(p+1)/2$ and the number of free parameters in $\boldsymbol{\phi}$ is $q+u+(p-u)u+u(u+1)/2+(p-u)(p-u+1)/2 = q+u+p(p+1)/2$, the envelope model reduces the total number of parameters by $p-u$.

Proposition 5.5. *Assume that for $i = 1, \ldots, n$, the predictors $\mathbf{X}_i$ are independent copies of a normal random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and variance $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$, and that the data $(Y_i, \mathbf{X}_i)$ are independent copies of $(Y, \mathbf{X})$ with finite fourth moments. Assume also that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal parameter vectors. Then, as $n \to \infty$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{env}} - \boldsymbol{\beta}_t)$ converges to a normal vector with mean 0 and covariance matrix*

$$\begin{aligned} \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) &= \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}) + \operatorname{avar}(\sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}}) \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T) \\ &\leq \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}), \end{aligned}$$

where $\mathbf{M} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_{u(p-u)}$, $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_0$.

The conditional log-likelihood is reflected in $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})$ primarily through the asymptotic variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, while $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ stem from the normal marginal likelihood of $\mathbf{X}$. The span of $\boldsymbol{\Gamma}$ reduces both $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, so the envelope serves as a link between the conditional and marginal likelihoods in the asymptotic variance. The other interpretations of this asymptotic result are similar to those given by Cook et al. (2013), where $\boldsymbol{\beta}$ is a vector of linear regression coefficients. The first part $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ is the same as in Proposition 5.2, which is the asymptotic covariance of estimator given the envelope. The second part $\operatorname{avar}(\sqrt{n}\mathbf{Q}_{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})$ reflects the asymptotic cost of estimating the envelope and this term is orthogonal to the first. Moreover, the envelope estimator is always more efficient than or equally efficient as the usual estimator $\widehat{\boldsymbol{\beta}}$.

An important special case of Proposition 5.5 is given in the following corollary.

Corollary 5.2. *Under the same conditions as in Proposition 5.5, if we assume further that $\boldsymbol{\Sigma}_{\mathbf{X}} = \sigma^2\mathbf{I}_p$, then $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \operatorname{span}(\boldsymbol{\beta}_t)$ and $\operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) = \operatorname{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$.*

Corollary 5.2 tells us that, if we have normal predictors with isotropic covariance, then the envelope estimator is asymptotically equivalent to the standard estimator and enveloping offers no gain, but there is no loss, either. This implies that there must be some degree of co-linearity among the predictors before envelopes can offer gains. We illustrate this conclusion in the simulations of Section 5.7.1.

Experience has shown that (5.4.4) provides a useful envelope estimator when the predictors satisfy the linearity condition but are not multivariate normal. In this case there is a connection between the desired envelope $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ and the marginal distribution of $\mathbf{X}$, as shown in Proposition 5.4, and $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is still a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$. If

the linearity condition is substantially violated, we can still use the objective function $\{C_n(\boldsymbol{\alpha},\boldsymbol{\beta}) + M_n(\boldsymbol{\Gamma})\}$ to estimate $\boldsymbol{\beta}$ within $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$ but this envelope may no longer equal $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$. Nevertheless, as demonstrated in Corollary 5.1, this has the potential to yield an estimator of $\boldsymbol{\beta}$ with smaller asymptotic variance than $\widehat{\boldsymbol{\beta}}$, although further work is required to characterize the gains in this setting. Alternatively, the 1D algorithm yields $\sqrt{n}$-consistent estimators without requiring the linearity condition.

5.5 Regression applications

5.5.1 Envelopes for weighted least squares

Consider a heteroscedastic linear model with data consisting of n independent copies of $(Y, \mathbf{X}, W)$, where $Y \in \mathbb{R}^1$, $\mathbf{X} \in \mathbb{R}^p$, and $W > 0$ is a weight with $\mathrm{E}(W) = 1$:

$$Y = \mu + \boldsymbol{\beta}^T\mathbf{X} + \varepsilon/\sqrt{W}, \tag{5.5.1}$$

where $\varepsilon \perp\!\!\!\perp (\mathbf{X}, W)$ and $\mathrm{var}(\varepsilon) = \sigma^2$. The constraint $\mathrm{E}(W) = 1$ is without loss of generality and serves to normalize the weights so they have mean 1 and to make subsequent expressions a bit simpler. Fitting under this model is typically done by using weighted least squares (WLS):

$$(\widehat{\mu}, \widehat{\boldsymbol{\beta}}) = \arg\min n^{-1}\sum_{i=1}^{n} W_i(Y_i - \mu - \boldsymbol{\beta}^T\mathbf{X}_i)^2, \tag{5.5.2}$$

where we normalize the sample weights so that $\overline{W} = 1$. Let $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{E}\{W(\mathbf{X} - \mathrm{E}(W\mathbf{X}))(\mathbf{X} - \mathrm{E}(W\mathbf{X}))^T\}$ be the weighted covariance matrix of $\mathbf{X}$, and let $\boldsymbol{\Sigma}_{\mathbf{X}Y(W)} = \mathrm{E}[W\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\}\{Y - \mathrm{E}(WY)\}]$ be the weighted covariance between $\mathbf{X}$ and Y. Then $\boldsymbol{\beta}_t = \boldsymbol{\Sigma}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Sigma}_{\mathbf{X}Y(W)}$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ converges to a normal random vector with mean 0 and variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2\boldsymbol{\Sigma}^{-1}_{\mathbf{X}(W)}$. According to Definition 5.1, we should now strive to estimate the $\sigma^2\boldsymbol{\Sigma}^{-1}_{\mathbf{X}(W)}$-envelope of $\mathrm{span}(\boldsymbol{\beta}_t)$, which can be done for a specified dimension u by using the 1D algorithm with $\widehat{\mathbf{M}} = s^2\mathbf{S}^{-1}_{\mathbf{X}(W)}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$, where s^2 is the estimator of σ^2 from the WLS fit (5.5.2), and $\mathbf{S}_{\mathbf{X}(W)}$ is the sample version of $\boldsymbol{\Sigma}_{\mathbf{X}(W)}$. Once an estimated basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\sigma^2\boldsymbol{\Sigma}^{-1}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t)$ is obtained, we re-fit the WLS regression model, replacing $\mathbf{X}$ with $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}$, to estimate the coordinate vector $\boldsymbol{\eta}$, :

$$(\widehat{\mu}, \widehat{\boldsymbol{\eta}}) = \arg\min n^{-1}\sum_{i=1}^{n} W_i(Y_i - \mu - \boldsymbol{\eta}^T\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}_i)^2. \tag{5.5.3}$$

The resulting envelope estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X}(W)}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X}Y(W)} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}(W)})}\widehat{\boldsymbol{\beta}}. \tag{5.5.4}$$

Cross validation or a hold out sample could now be used straightforwardly as an aid to selecting u. When the weights are constant, this method has the same structure as partial

least squares regression with the important exception that partial least squares uses the SIMPLS algorithm in place of the 1D algorithm that we recommend.

To take a likelihood approach with the envelope $\mathcal{E}_{\sigma^2\mathbf{\Sigma}^{-1}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t)$, we follow Section 5.4.1 to decompose the full log-likelihood into the sum of the conditional log-likelihood from $Y|(\mathbf{X}, W)$, denoted by $C_n(\mu, \boldsymbol{\beta}, \sigma^2) \equiv C_n(\mu, \mathbf{\Gamma}, \boldsymbol{\eta}, \sigma^2)$, and the conditional log-likelihood from $\mathbf{X}|W$, denoted by $M_n(\boldsymbol{\psi}_1, \mathbf{\Gamma})$, where $\boldsymbol{\psi}_1$ denotes additional nuisance parameters. We condition on the observed values of W, assuming that it is ancillary. Then holding $\mathbf{\Gamma}$ fixed and partially maximizing $C_n(\mu, \mathbf{\Gamma}, \boldsymbol{\eta}, \sigma^2) + M_n(\boldsymbol{\psi}_1, \mathbf{\Gamma})$, we get a likelihood-based objective function $L_n(\mathbf{\Gamma})$ that plays the same role as the moment-based objective function $J_n(\mathbf{\Gamma})$. Once an basis $\widehat{\mathbf{\Gamma}}$ is obtained by maximizing $L_n(\mathbf{\Gamma})$, the final estimators follow from (5.5.3) and (5.5.4).

Two kinds of assumptions on the distribution of $\mathbf{X}|W$ could be made depending on the sampling model. One is to assume that the predictors and the weights are independently sampled and that the predictor vector follows a multivariate normal distribution. By Proposition 5.4, we have $\mathcal{E}_{\sigma^2\mathbf{\Sigma}^{-1}_{\mathbf{X}(W)}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$ and the partially maximized log-likelihood up to a constant is (see derivation in the Supplement):

$$L_n(\mathbf{\Gamma}) \;=\; -\frac{n}{2}\left\{\log(s^2_{\mathbf{\Gamma}}) + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\mathbf{\Gamma}|\right\}, \tag{5.5.5}$$

where $s^2_{\mathbf{\Gamma}} = s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)}$ is the estimator of σ^2 from (5.5.3) with fixed $\mathbf{\Gamma}$ and $s^2_{Y(W)} = \sum_{i=1}^n W_i(Y_i - \sum_{j=1}^n W_jY_j/n)/n$ is the sample weighted variance for Y.

Alternatively, it might be reasonable to assume that $\mathbf{X}|W$ has a weighed variance similar to the error variance from $Y|(\mathbf{X}, W)$: $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\mathbf{\Sigma})$. Then the partially maximized log-likelihood up to a constant is (see derivation in Section 5.9.8):

$$\begin{aligned} L_n(\mathbf{\Gamma}) \;&=\; -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T s^2\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T(s^2\mathbf{S}^{-1}_{\mathbf{X}(W)} + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T)^{-1}\mathbf{\Gamma}|\right\} \\ &=\; -\frac{n}{2}\left\{\log|\mathbf{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}|Y(W)}\mathbf{\Gamma}|\right\}, \end{aligned} \tag{5.5.6}$$

where $\mathbf{S}_{\mathbf{X}|Y(W)} = \mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)}\mathbf{S}^T_{\mathbf{X}Y(W)}/s^2_{Y(W)}$ and the first equality shows equivalence between maximizing this likelihood-based objective function of $\mathbf{\Gamma}$ and minimizing the moment based objective function $J_n(\mathbf{\Gamma})$ with $\widehat{\mathbf{M}} = s^2\mathbf{S}^{-1}_{\mathbf{X}(W)}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$.

The asymptotic results in Proposition 5.5 hold for $\widehat{\boldsymbol{\beta}}_{\text{env}}$ if (i) $\mathbf{X}$ is normal and independent of W, and we use (5.5.5) to get $\widehat{\mathbf{\Gamma}}$; or (ii) $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\mathbf{\Sigma})$ and we use (5.5.6) to get $\widehat{\mathbf{\Gamma}}$. Then the asymptotic covariance of the envelope estimator will be asymptotically less than or equal to that of the standard WLS estimator. Moreover, as long as the model (5.5.1) holds and both ε and $\mathbf{X}|W$ follow a distribution with finite fourth moments, asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is guaranteed no matter which $L_n(\mathbf{\Gamma})$ we choose to maximize. Finally, if we fix $W = 1$, then both objective functions in (5.5.5) and in (5.5.6) reduces

to (5.2.1), which corresponds to the envelope model for reducing $\mathbf{X}$ in the homoscedastic linear models (Cook et al. 2013).

5.5.2 Generalized linear models with canonical link

In the textbook generalized linear model (GLM) setting (Agresti 2002), Y belongs to an exponential family with probability mass or density function $f(Y_i|\vartheta_i,\varphi) = \exp\{[Y_i\vartheta_i - b(\vartheta_i)]/a(\varphi) + c(Y,\varphi)\}$, $i = 1,\ldots,n$, where ϑ is the natural parameter and $\varphi > 0$ is the dispersion parameter. We consider the canonical link function $\vartheta(\alpha,\boldsymbol{\beta}) = \alpha + \boldsymbol{\beta}^T\mathbf{X}$, which is a monotonic differentiable function of $\mathrm{E}(Y|\mathbf{X},\vartheta,\varphi)$. And $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter vector of interest. We also restrict discussion to one-parameter families such as binomial, Poisson and exponential regressions, so that the dispersion parameter φ is not needed. For two-parameter families, such as normal, Gamma and inverse Gamma regressions, φ is a nuisance parameter and our proposed envelope methods for efficient estimation of $\boldsymbol{\beta}$ can be developed similarly.

Consider moment-based estimation based on Definition 5.1 and the 1D algorithm. The conditional log likelihood takes the form of $\log f(y|\vartheta) = y\vartheta - b(\vartheta) + c(y) \equiv \mathcal{C}(\vartheta)$, where $\vartheta = \alpha + \boldsymbol{\beta}^T\mathbf{X}$ is the canonical parameter. Then the Fisher information matrix for $(\alpha,\boldsymbol{\beta})$ evaluated at the true parameters is

$$\mathbf{F}(\alpha_t,\boldsymbol{\beta}_t) = \mathrm{E}\left(-\mathcal{C}''(\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X})\begin{pmatrix} 1 & \mathbf{X} \\ \mathbf{X}^T & \mathbf{X}\mathbf{X}^T \end{pmatrix}\right), \tag{5.5.7}$$

where $\mathcal{C}''(\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X})$ is the second derivative of $\mathcal{C}(\vartheta)$ evaluated at $\alpha_t + \boldsymbol{\beta}_t^T\mathbf{X}$. To induce orthogonal parameters $(\alpha,\boldsymbol{\beta}) \mapsto (a,\boldsymbol{\beta})$ for direct application of Proposition 5.5, we define the weights $W(\vartheta) = \mathcal{C}''(\vartheta)/\mathrm{E}(\mathcal{C}''(\vartheta))$ so that $\mathrm{E}(W) = 1$, and write $\vartheta = \alpha + \boldsymbol{\beta}^T\mathrm{E}(W\mathbf{X}) + \boldsymbol{\beta}^T\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\} := a + \boldsymbol{\beta}^T\{\mathbf{X} - \mathrm{E}(W\mathbf{X})\}$. Then the new parameterization $(a,\boldsymbol{\beta})$ has Fisher information matrix

$$\mathbf{F}(a_t,\boldsymbol{\beta}_t) \;=\; \mathrm{E}(-\mathcal{C}'')\begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathbf{X}(W)} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = \mathrm{E}\{W[\mathbf{X} - \mathrm{E}(W\mathbf{X})][\mathbf{X} - \mathrm{E}(W\mathbf{X})]^T\}$. We now have orthogonal parameters and $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}(a_t,\boldsymbol{\beta}_t) = \left\{\mathrm{E}(-\mathcal{C}'')\cdot\boldsymbol{\Sigma}_{\mathbf{X}(W)}\right\}^{-1}$, while $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t)$ is the corresponding envelope. Using parameterization $(a,\boldsymbol{\beta})$ or $(\alpha,\boldsymbol{\beta})$ lead to equivalent implementation since we are interested only in $\boldsymbol{\beta}$. Therefore, we stay with the usual parameter $(\alpha,\boldsymbol{\beta})$ in future discussion.

By Definition 5.1, this leads us to the 1D estimator with $\widehat{\mathbf{M}} = \mathbf{S}^{-1}_{\mathbf{X}(W)}/\left\{n^{-1}\sum_{i=1}^n(-\mathcal{C}''(\widehat{\vartheta}_i))\right\}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$, where $\widehat{\vartheta}_i = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}^T\mathbf{X}_i$ is the estimated canonical link function and $\mathbf{S}_{\mathbf{X}(W)}$ is the sample estimator of $\boldsymbol{\Sigma}_{\mathbf{X}(W)}$. However, the weight $W(\vartheta)$ depends on the parameter $\boldsymbol{\beta}$ so that iterative updates of weights and estimators could used to refine the final estimator,

as discussed at the end of Section 5.3. Cross validation or a hold-out sample could again be used as an aid to selecting the dimension of the envelope.

Turning to a likelihood approach based on normal predictors, it follows from Section 5.4.1 that the full log-likelihood can be written as $L_n(\alpha, \boldsymbol{\Gamma}, \boldsymbol{\eta}) = C_n(\alpha, \boldsymbol{\beta}) + M_n(\boldsymbol{\Gamma})$, where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ is the coefficient vector of interest and $M_n(\boldsymbol{\Gamma})$ is given in Lemma 5.1. The conditional log-likelihood, which varies for different exponential family distribution of $Y|\mathbf{X}$, can be written as $C_n(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n \mathcal{C}(\vartheta_i)$, where $\mathcal{C}(\vartheta_i)$ is summarized in Table 5.1.

	$\mu \equiv \mathrm{E}(Y\|\mathbf{X}, \alpha, \boldsymbol{\beta})$	$\mathcal{C}(\vartheta)$	$\mathcal{C}'(\vartheta)$	$-\mathcal{C}''(\vartheta)$ $(\propto W)$
Normal	ϑ	$Y\vartheta - \vartheta^2/2$	$Y - \vartheta$	1
Poisson	$\exp(\vartheta)$	$Y\vartheta - \exp(\vartheta)$	$Y - \exp(\vartheta)$	$\exp(\vartheta)$
Logistic	$\exp(\vartheta)/A(\vartheta)$	$Y\vartheta - \log A(\vartheta)$	$Y - \exp(\vartheta)/A(\vartheta)$	$\exp(\vartheta)/A^2(\vartheta)$
Exponential	$-\vartheta^{-1} > 0$	$Y\vartheta - \log(-\vartheta)$	$(Y - 1/\vartheta)$	ϑ^{-2}

Table 5.1: A summary for the mean functions, the conditional log-likelihoods and their derivatives of various exponential family distributions. $A(\vartheta) = 1 + \exp(\vartheta)$.

Fisher scoring is the usual iterative method for fitting based on $C_n(\alpha, \boldsymbol{\beta})$ and its update procedure can be summarized in the form of a weighted least squares (WLS) estimator as follows:

$$\boldsymbol{\beta}^{(k+1)} = \mathbf{S}^{-1}_{\mathbf{X}(W^{(k)})}\mathbf{S}_{\mathbf{X}V^{(k)}(W^{(k)})}, \tag{5.5.8}$$

where $W^{(k)} = W(\vartheta^{(k)}) = W(\alpha^{(k)} + \mathbf{X}^T\boldsymbol{\beta}^{(k)})$ is the weight at the k-th iteration, $V^{(k)} = \vartheta^{(k)} + \left[Y - \mu(\vartheta^{(k)})\right]/W^{(k)}$ is a pseudo-response variable at the k-th iteration and the weighted covariance $\mathbf{S}_{\mathbf{X}V^{(k)}(W^{(k)})}$ is defined in the same way as $\mathbf{S}_{\mathbf{X}(W)}$. The Fisher scoring process stops when $|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}|$ is less than a specified tolerance, we then obtain the estimator $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(k)}$ and estimators $\widehat{\alpha}$, $\widehat{\vartheta}$, $\widehat{W}$ and $\widehat{V}$ at step k. Upon convergence of the Fisher scoring procedure, the WLS formulation in (5.5.8) simplifies to a version of the moment-based objective function $J_n(\boldsymbol{\Gamma})$ with $\widehat{\mathbf{M}} = \mathbf{S}^{-1}_{\mathbf{X}(W)}/\left\{n^{-1}\sum_{i=1}^n(-\mathcal{C}''(\widehat{\vartheta}_i))\right\}$ and $\widehat{\mathbf{U}} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T$:

$$J_n(\boldsymbol{\Gamma}) = \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(\widehat{W})}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}|\widehat{V}(\widehat{W})}\boldsymbol{\Gamma}|, \tag{5.5.9}$$

where $\mathbf{S}_{\mathbf{X}|\widehat{V}(\widehat{W})} = \mathbf{S}_{\mathbf{X}|\widehat{V}(\widehat{W})} = \mathbf{S}_{\mathbf{X}(\widehat{W})} - \mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}\mathbf{S}^T_{\mathbf{X}\widehat{V}(\widehat{W})}/s^2_{\widehat{V}(\widehat{W})}$ is a weighted residual covariance matrix defined in the same way as $\mathbf{S}_{\mathbf{X}|Y(W)}$ in WLS envelope estimation of Section 5.5.1, equation (5.5.6).

The WLS formulation in (5.5.8) also facilitates the partial maximization of $C_n(\alpha, \boldsymbol{\beta}) = C_n(\alpha, \boldsymbol{\eta}, \boldsymbol{\Gamma})$ over α and $\boldsymbol{\eta}$. When $\boldsymbol{\Gamma}$ is fixed, we have $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} = (\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(\widehat{W})}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}\widehat{V}(\widehat{W})}$. Therefore, the partially maximized log-likelihood for $\boldsymbol{\Gamma}$ at convergence of the Fisher scoring

procedure is

$$\begin{aligned} L_n(\boldsymbol{\Gamma}) &= C_n(\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma}) \\ &= C_n(\widehat{\alpha}, \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}) - \frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}. \quad (5.5.10) \end{aligned}$$

The iterative update between $\widehat{\boldsymbol{\Gamma}}$ and other estimators is then incorporated with the Fisher scoring methods as follows. Let the standard estimators be initial values for the following iterative update and be denoted by $\widehat{\boldsymbol{\beta}}^{(0)}$, $\widehat{\alpha}^{(0)}$, $\widehat{\vartheta}^{(0)}$, $\widehat{W}^{(0)}$ and $\widehat{V}^{(0)}$. Then for $k = 0, 1, \ldots,$

1. update $\widehat{\boldsymbol{\Gamma}}^{(k+1)}$ by maximizing (5.5.10) over Grassmannian $\mathcal{G}_{p,u}$, where $C_n(\boldsymbol{\Gamma})$ is calculated based on $\widehat{a}^{(k)}$, $\widehat{\boldsymbol{\beta}}^{(k)}$, $\widehat{\vartheta}^{(k)}$, $\widehat{W}^{(k)}$ and $\widehat{V}^{(k)}$.

2. update $\widehat{a}^{(k+1)}$ and $\widehat{\boldsymbol{\eta}}^{(k+1)}$ by using Fisher scoring method to fit GLM of Y on $(\widehat{\boldsymbol{\Gamma}}^{(k+1)})^T\mathbf{X}$. Then let $\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\Gamma}}^{(k+1)}\widehat{\boldsymbol{\eta}}^{(k+1)}$ and simultaneously update $\widehat{\vartheta}^{(k+1)}$, $\widehat{W}^{(k+1)}$ and $\widehat{V}^{(k+1)}$.

We solve Step 1 by using the *sg_min* Matlab package by Ross A. Lippert (`http://web.mit.edu/~ripper/www/software/`). To implement Step 1 using *sg_min* package, we also need to compute the matrix derivative $\partial L_n(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma}$, which is included in the Supplement. Also, Step 1 can be replaced by minimizing (5.5.9) using the 1D algorithm. The Matlab package *glmfit* will take care of Step 2 using the Fisher scoring method.

5.5.3 Envelopes for Cox regression

In the study of survival time T on the p-dimensional covariates $\mathbf{Z} \in \mathbb{R}^p$, Cox's proportional hazards model (Cox 1972, 1975) assumes the hazard function $h(t|\mathbf{Z})$ of a subject with covariates $\mathbf{Z}$ has the form

$$h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}), \quad (5.5.11)$$

where $h_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients in which we are primarily interested. Let X_i and C_i be the failure time and the censoring time of the i-th subject, $i = 1, \ldots, n$. Define $\delta_i = I(X_i \le C_i)$ and $T_i = \min(X_i, C_i)$. The data then consists of $(T_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. The failure time and the censoring time are assumed to be independent given the covariates. We assume that there is no ties for the observed times.

Continuing from the discussion of Section 5.5.3, the log partial likelihood for $\boldsymbol{\beta}$ is

$$C_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\delta_i\left\{\boldsymbol{\beta}^T\mathbf{Z}_i - \log\left(\sum_{j=1}^{n} I(T_j \ge T_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\right)\right\}, \quad (5.5.12)$$

whose first-order derivative (score function) and second-order derivative (Hessian matrix) are

$$\begin{aligned} C_n'(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \delta_i \left\{ \mathbf{Z}_i - \sum_{j=1}^{n} W_{ij} \mathbf{Z}_j \right\}, \\ C_n''(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \delta_i \left\{ \left(\sum_{j=1}^{n} W_{ij} \mathbf{Z}_j \mathbf{Z}_j^T \right) - \left(\sum_{j=1}^{n} W_{ij} \mathbf{Z}_j \right) \left(\sum_{j=1}^{n} W_{ij} \mathbf{Z}_j^T \right) \right\}, \end{aligned}$$

where $W_{ij} = I(T_j \geq T_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) / \left\{ \sum_{j=1}^{n} I(T_j \geq T_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) \right\}$ is the weight with respect to subject i. The Fisher information matrix is $\mathbf{J}_n(\boldsymbol{\beta}) = -\lim_{n\to\infty} n^{-1} C_n''(\boldsymbol{\beta})$. The implementation of envelope Cox model is similar to that of envelope GLM in Section 5.5.2 by adding the same $M_n(\boldsymbol{\Gamma})$ to the objective function. Theoretical properties of the envelope Cox model is included in Section 5.4.1.

5.5.4 Other regression applications

Based on our general framework, the envelope model can be adapted to numerous other studies. For instances, in Supplement Sections 5.5.1 and 5.9.8, we give derivation and implementation details for envelope in weighted least squares regression; in Supplement Section 5.5.3, we include details for envelopes in Cox regression. Theoretical results in this section could also apply to the WLS and the Cox regression models, as we discussed in the Supplement Section 5.5.

The envelope methods proposed here are based on sample estimators of asymptotic covariance matrices, which may be sensitive to outliers. However, the idea of envelopes can be extended to robust estimation procedures (see, for example, Yohai et al. (1991)).

5.6 Simulations

In this section we present a few simulations to support and illustrate the foundations discussed in previous sections. For each of the simulation settings, we simulated 100 datasets for each of the three sample sizes: $n = 50$, 200 and 800. The true dimension of the envelope was always used in estimation.

5.7 Least squares regression

In this section, we report the results of numerical studies on the robustness of three estimators in linear regression: (1) the ordinary least squares (OLS) estimator, (2) the PLS estimator from the SIMPLS algorithm, and (3) the moment-based envelope estimator from the 1D algorithm in Cook and Zhang (2014a). Same as in Section 5.6, for each of the

t_6	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	57	34	6	85	44	48
S.E.	1.6	0.08	0.003	0.6	1.2	2.7
$n = 200$	40	32	3	80	35	14
S.E.	1.8	0.8	0.09	1.0	0.8	0.6
$n = 800$	23	27	1	71	30	7
S.E.	1.4	1.0	0.5	1.5	0.8	0.2

$U(0,1)$	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	27	33	1.8	74	33	12.6
S.E.	2	1	0.05	2	1	0.5
$n = 200$	14	29	0.8	59	31	5.4
S.E.	1	1	0.03	2	1	0.2
$n = 800$	6.8	24	0.4	44	25	2.8
S.E.	0.4	1	0.01	2	1	0.1

χ^2_4	$\mathbf{\Sigma_X}$			$0.01\mathbf{\Sigma_X}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	73	35	13	88	60	75
S.E.	1.4	1.0	0.8	0.3	2	2
$n = 200$	57	33	6.0	85	47	52
S.E.	2	1	0.2	1	1	2
$n = 800$	38	27	2.9	79	35	13
S.E.	2	1	0.1	1	1	1

Table 5.2: Simulations for heterogeneous covariance matrix $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. Averaged angles between the true parameter vector and three estimators for the least squares simulation. Standard error is included below each value.

t_6		Standard	SIMPLS	1D
$n = 50$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	27.8 (0.7)	33.3 (0.7)	30.9 (0.9)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.32	0.79	1.22
$n = 200$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	13.1 (0.3)	18.1 (0.4)	14.0 (0.4)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.06	0.91	1.05
$n = 800$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	6.6 (0.2)	9.0 (0.2)	6.7 (0.2)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.02	0.98	1.02

$U(0,1)$		Standard	SIMPLS	1D
$n = 50$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	6.8 (0.2)	22.2 (0.5)	6.9 (0.2)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.02	0.71	1.01
$n = 200$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	3.3 (0.08)	11.7 (0.3)	3.3 (0.08)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.00	0.89	1.00
$n = 800$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	1.7 (0.04)	5.9 (0.1)	1.7 (0.04)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.00	0.97	1.00

χ^2_4		Standard	SIMPLS	1D
$n = 50$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	51 (1.4)	50 (1.2)	53 (1.4)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.42	1.21	1.39
$n = 200$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	29.9 (0.8)	30.6 (0.7)	32.6 (0.9)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.39	1.16	1.28
$n = 800$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	15.0 (0.4)	16.2 (0.4)	15.9 (0.4)
	$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.08	1.04	1.07

Table 5.3: Simulations for isotropic covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = 0.01\mathbf{I}_p$. Averaged angles and length ratios between the true parameter vector and various estimators. Standard errors of the averaged angles are included in parentheses.

simulation settings, we simulated 100 datasets for each of the three sample sizes: $n = 50$, 200 and 800. The true dimension of the envelope was always used in estimation. We used the Matlab function *plsregress* to compute the PLS estimator. From Proposition 4.1 in Cook et al. (2013), we know that the SIMPLS algorithm produces $\sqrt{n}$-consistent envelope estimators when the number of component in the algorithm is chosen to be the dimension of the envelope. The univariate response was generated as $Y = \boldsymbol{\beta}^T\mathbf{X} + \varepsilon$, where the elements of ε were $U(0, 1)$ variates. The covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, was generated as it was for logistic and the Poisson regression, except the eigenvalues of $\boldsymbol{\Omega}$ were 1 and 5, while the eigenvalues of $\boldsymbol{\Omega}_0$ ranged from 0.0002 to 20. These choices introduce a degree of collinearity in the immaterial variation $\text{var}(\boldsymbol{\Gamma}_0^T\mathbf{X})$, which is the kind of regression that PLS estimators were designed to handle. We also used $0.01\boldsymbol{\Sigma}_{\mathbf{X}}$ instead of $\boldsymbol{\Sigma}_{\mathbf{X}}$ to compare performance with a weaker signal.

The angles are shown in Table 5.2, the lengths being excluded because they produced no comparative conclusion beyond those available from the results shown. The 1D algorithm did the best in all situations, except when a relatively small sample sizes was combined with a weak signal ($0.01\boldsymbol{\Sigma}_{\mathbf{X}}$) and t_6 or χ^2_4 errors. In those situations the SIMPLS might have a slight edge. Because there was strong collinearity in $\mathbf{X}$, OLS did notably worse than the other two methods, as expected. One exception to this was when $n = 800$ with uniformly distributed predictors.

We also repeated the simulations of Table 5.2 when $\boldsymbol{\Sigma}_{\mathbf{X}} = 0.01\mathbf{I}_p$, so envelope methods reduce to the standard OLS estimator. The OLS and 1D estimators performed essentially the same in all cases, while the SIMPLS algorithm did a bit worse with t_6 and χ^2_4 errors, and notably worse with $U(0, 1)$ errors. Table 5.3 gives the results of the least squares simulation with predictor covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = 0.01\mathbf{I}_p$.

5.7.1 Generalized linear models

In this section we report the result of numerical studies to compare three estimators in the contexts of logistic and Poisson regression: (1) the standard GLM estimators obtained by the Fisher scoring method, (2) iteratively re-weighted partial least squares estimators for GLMs (IRPLS; Marx 1996), and (3) envelope estimators based on minimizing the likelihood-based objective function (5.5.10) over the appropriate Grassmannian. The 1D algorithm (Algorithm 1) was used to provide $\sqrt{n}$-consistent starting values for Grassmannian optimizations. We used the Matlab function *glmfit* to obtain the standard GLM estimators. The IRPLS algorithm is a widely recognized extension of PLS algorithms for GLMs. It embeds a weighted partial least squares algorithm in the Fisher scoring method and iteratively updates between the reduced predictors $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}$ and the other parameters $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\vartheta}, \widehat{V}, \widehat{W})$, similar to our envelope method described in Section 5.5.2. The original IRPLS algorithm by Marx (1996) is based on the NIPALS algorithm (Wold 1966) for PLS,

but our simulations all used the SIMPLS algorithm (de Jong 1993) within IRPLS. This is because SIMPLS is more widely used and is implemented in the *plsregress* function in Matlab.

In all these simulations, the envelope dimension was equal to 2, $p = 10$ and $\mathbf{X} \sim N(0, \mathbf{\Sigma_X})$, where $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, $\mathbf{\Gamma} \in \mathbb{R}^{p\times 2}$ was generated by filling with independent uniform $(0, 1)$ variates and then standardized so that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ is an orthogonal matrix. The positive definite symmetric matrices $\mathbf{\Omega} \in \mathbb{R}^{2\times 2}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{8\times 8}$ were generated as $\mathbf{ODO}^T$ where $\mathbf{O}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix of positive eigenvalues. We chose the eigenvalues of $\mathbf{\Omega}$ to be 0.1 and 0.5, while the eigenvalues of $\mathbf{\Omega}_0$ range from 0.002 to 20. The true parameter vector $\boldsymbol{\beta}_t = \mathbf{\Gamma}\boldsymbol{\eta}_t$ with $\boldsymbol{\eta}_t = (1, 1)^T$. The intercepts were $\alpha = 2$ for Poisson regression and $\alpha = 1$ for logistic regression.

The averaged angles $\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$ and length ratios $||\widehat{\boldsymbol{\beta}}||/||\boldsymbol{\beta}_t||$ are summarized in Table 5.4. From this table, we found that the envelope estimator was always the best for various sample sizes. The advantages of envelope estimator over its competitors were clear in both logistic and Poisson regressions. The improvements of envelope estimators over the standard estimators became more substantial with increased sample sizes. At the same time, the standard estimator always did better than the IRPLS estimator in terms of angle.

Recall that, according to Corollary 5.2, we need some degree of co-linearity among the predictors before envelope can offer gains. To illustrate this conclusion, we repeated the sets of simulations by using $\mathbf{\Sigma_X} = \mathbf{I}_p$ for Poisson regression and $\mathbf{\Sigma_X} = 4\mathbf{I}_p$ for logistic regression. From the results summarized in Table 5.5, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\text{env}}$ has similar performance, as expected.

5.7.2 Cox regression

In the study of survival time T on the p-dimensional covariates $\mathbf{Z} \in \mathbb{R}^p$, Cox's proportional hazards model (Cox 1972, 1975) assumes the hazard function $h(t|\mathbf{Z})$ of a subject with covariates $\mathbf{Z}$ has the form

$$h(t|\mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}), \tag{5.7.1}$$

where $h_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients in which we are primarily interested. Let X_i and C_i be the failure time and the censoring time of the i-th subject, $i = 1, \ldots, n$. Define $\delta_i = I(X_i \le C_i)$ and $T_i = \min(X_i, C_i)$. The data then consists of $(T_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. The failure time and the censoring time are assumed to be independent given the covariates. We assume that there is no ties for the observed times.

In our setting, $\mathbf{Z} \in \mathbb{R}^{10}$ followed the multivariate normal distribution $N(0, \mathbf{\Sigma_Z})$ where

$$\mathbf{\Sigma_Z} = \mathbf{\Gamma\Omega\Gamma}^T + 0.2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T,$$

			Standard	IRPLS	Envelope
Poisson	$n = 50$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	38 (1.5)	55 (0.8)	**25 (2.0)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.4	0.21	1.4
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	21 (1.1)	48 (0.6)	**7 (0.8)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.09	0.25	1.03
	$n = 800$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	9 (0.5)	46 (0.4)	**2 (0.3)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.02	0.28	1.01
Logistic	$n = 50$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	75 (1)	84 (1)	**74 (2)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	10.2	0.22	8.1
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	62 (2)	81 (1)	**36 (3)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	3.0	0.12	2.0
	$n = 800$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	45 (2)	77 (0.4)	**9 (1)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.6	0.087	1.0

Table 5.4: Simulations with $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. We summarize averaged angles and length ratios between the true parameter vector and various estimators. The best performances are marked in bold. Standard errors are included in the parentheses. Standard errors for the length ratios are less than or equal to 0.01.

			Standard	IRPLS	Envelope
Poisson	$n = 50$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	8.9 (0.3)	46 (0.8)	9.2 (0.9)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.02	0.58	1.02
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	3.7 (0.1)	38 (0.5)	3.7 (0.1)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.00	0.70	1.00
	$n = 800$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	1.74 (0.04)	36 (0.3)	1.74 (0.04)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.00	0.77	1.00
Logistic	$n = 50$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	12.6 (1.6)	17.9 (2.3)	13.7 (1.8)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.18	0.27	1.09
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	16.2 (0.4)	29.2 (0.7)	16.9 (0.4)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.15	0.79	1.14
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \hat{\boldsymbol{\beta}})$	8.0 (0.2)	17.3 (0.4)	8.1 (0.2)
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.03	0.90	1.03

Table 5.5: Simulations for isotropic covariance matrices: $\boldsymbol{\Sigma}_{\mathbf{X}} = 4\mathbf{I}_p$ for logistic regression and $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}_p$ for Poisson regression. We summarized averaged angles and length ratios between the true parameter vector and various estimators. Standard errors are included in the parentheses. Standard errors for the length ratio are less than or equal to 0.01.

			Standard	PLS	Envelope
Cox	$n = 50$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	50 (0.9)	23 (0.8)	**15 (1)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.95	0.99	0.97
	$n = 200$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	25 (0.6)	15 (0.5)	**6 (0.2)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.16	1.00	1.00
	$n = 800$	$\angle(\boldsymbol{\beta}_t, \widehat{\boldsymbol{\beta}})$	13 (0.3)	10 (0.3)	**3 (0.07)**
		$\|\|\widehat{\boldsymbol{\beta}}\|\|/\|\|\boldsymbol{\beta}_t\|\|$	1.03	1.00	1.00

Table 5.6: Cox regression: Averaged angles and length ratios between the true parameter vector and various estimators. The best performances are marked in bold. Standard errors are included in parentheses. Standard errors for the length ratio are less than or equal to 0.01.

where $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Omega}$ were generated in the same way as in Section 5.7.1, except that eigenvalues of $\boldsymbol{\Omega}$ were 1 and 5. The true parameter vector $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_t$ where $\boldsymbol{\eta}_t = (1, 1)^T$. Then we followed the simulation model in Nygard and Borgan (2008), where the survival time followed a Weibull distribution with scale parameter $\exp(\boldsymbol{\beta}^T\mathbf{Z}/5)$ and shape parameter 5, which was a Weibull distribution with hazard rate $h(Y|\mathbf{Z}) = 5Y^4 \cdot \exp(\boldsymbol{\beta}^T\mathbf{Z})$. The censoring variable δ was generated as the integer part of a uniform$(0, 2)$ random variable, which gave censoring rates of approximately 50%.

We considered three different estimators: the standard estimator without envelope structure, the likelihood-based envelope estimator (see Supplement Section 5.5.3 for implementation details), and partial least square for Cox regression (Nygard and Borgan 2008). From Table 5.6, we found that the envelope estimator was always the best, while the PLS estimator also improved over the standard estimator.

5.8 Illustrative data analysis

We present four small examples in this section to illustrate selected aspects of the proposed foundations.

5.8.1 Logistic regression: Australian Institute of Sport data

This data set, originally from Cook and Weisberg (1994; page 98), contains various measurements on 102 male and 100 female athletes collected at the Australian Institute of Sport. We use this example to illustrate the phenomenon in Section 5.2.3. We take height (cm) to be X_1, weight (kg) to be X_2 and Y as a binary indicator for male or female athletes. A plot of the data, Figure 5.2, is similar to that in Figure 5.1(a) so we expected some gain from envelopes. The standard estimator $\widehat{\boldsymbol{\beta}} = (0.123, 0.062)^T$ with standard

errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}) = (0.034, 0.022)^T$, and envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = (0.085, 0.086)^T$ with standard errors $\mathrm{SE}(\widehat{\boldsymbol{\beta}}_{\mathrm{env}}) = (0.014, 0.014)^T$. The envelope estimator compared to the standard estimator, had 40% and 60% smaller standard errors for the two coefficients in $\boldsymbol{\beta}$.

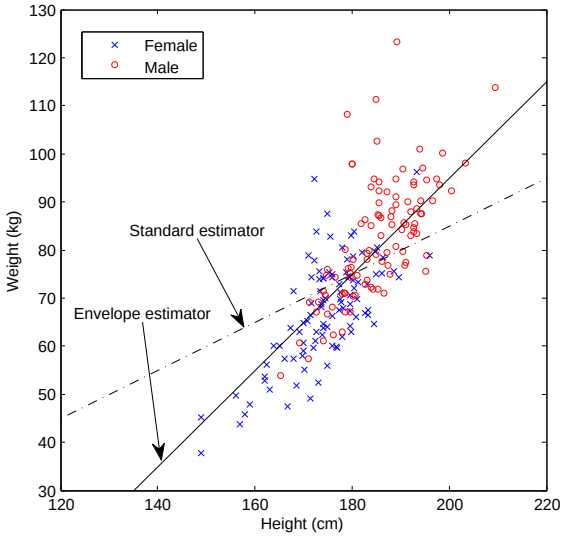


Figure 5.2: Heights and weights of male and female athletes.

5.8.2 Logistic regression: colon cancer diagnosis

We use these data to illustrate the classification power of the envelope estimator as well as its estimation efficiency. The objective of this study was to distinguish normal tissues and adenomas (precancerous tissues) that were found during colonoscopy. When the normal and adenomatous tissues are illuminated with ultraviolet light, they fluoresce at different wavelengths. The $p = 22$ predictors, which are laser-induced fluorescence (LIF) spectra measured at 8 nm spacing from 375 to 550 nm, reflect this phenomenon. The data set consists of $n = 285$ tissue samples, comprised of 180 normal tissues and 105 adenomas. Studies on a similar data set can be found in Hawkins and Maboudou-Tchao (2013).

The envelope dimension was chosen to be 2 by cross-validation. Compared to the standard logistic regression, the bootstrap standard errors of the 22 elements in $\widehat{\boldsymbol{\beta}}$ were 2.2 to 12.5 times larger than the bootstrap standard errors of the elements in $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$. On the other hand, five-fold cross-validation classification error rate for the standard logistic regression estimator was 21% while it was 17.5% for the envelope estimator with $u = 2$ or $u = 3$. And the error rate for envelope estimator with any dimension between 1 and 10 was no larger than that of the standard estimator. To gain some idea about the level of intrinsic (non-estimative) variation in the envelope error rate, we took $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ to be the true $\boldsymbol{\beta}$, sampled the predictors 1000 times and predicted the response using the true $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathrm{env}}$. The resulting classification error was 16.7%, which is not far from the observed error rate

of 17.5%.

5.8.3 Linear discriminant analysis: wheat protein data

In this data set (Cook et al. 2010), we have a binary response variable Y indicating high or low protein content and six highly correlated predictors which are the logarithms of near infrared reflectance at six wavelengths across the range 1680-2310nm. The correlations between these predictors range from 0.9118 to 0.9991. We standardized the predictors marginally so that each had mean 0 and variance 1. We illustrate the possibility of using envelopes to improve Fisher's linear discriminant analysis (LDA).

It has been shown (Ye 2007) that two-class LDA is equivalent to a least squares problem: no matter how we code Y for two classes, the least square fitted coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is always proportional to the LDA direction. Hence, we can use $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T\mathbf{X}$ as the discriminant direction and fit the LDA classification rule based on it. Proceeding similarly can replace the OLS coefficient vector with the coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ estimated by SIMPLS algorithm and the coefficient vector and $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ estimated by using the 1D algorithm to improve OLS. We estimated the error rates for these three classifiers as the average error rate over 100 replications of five-fold cross-validations on the $n = 50$ samples. The average error rates are shown in Table 5.7. Clearly the 1D algorithm had the best classification performance.

	Standard	PLS	Envelope
Error rate	10.2	57.1	5.4
S.E.	0.4	0.6	0.2

Table 5.7: Five-fold cross-validation error rates based on Fisher's LDA and expressed as percentages.

5.8.4 Poisson regression: horseshoe crab data

	Standard	IRPLS	Envelope
Pearson's χ^2	642.1	556.1	553.7
S.E.	7.8	1.6	1.6

Table 5.8: Performance of four methods applied to the horseshoe crab data.

This is a textbook Poisson regression dataset about nesting horseshoe crabs (see Agresti (2002), Section 4.3). The response is the number of satellite male crabs residing near a female crab. Explanatory variables included the female crab's color, spine condition,

weight, and carapace width. Since color is a factor with four levels, we use three indicator variables, X_1, X_2 and X_3, to indicate color. Also, spine condition is a factor with three levels, so we used two indicator variables X_4 and X_5 for spine condition. The remaining two continuous predictors are: X_6 the weight and X_7 the carapace width. The sample size is $n = 173$.

Five-fold cross-validation was used to determine the dimension of the envelope, giving $u = 1$, which was also the answer by BIC. To assess the gain in envelope reduction with $u = 1$, we repeated the five-fold cross-validation procedure 100 times and obtained the averaged Pearson's χ^2 statistic $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2/\widehat{Y}_i$ and its standard error shown in Table 5.8. Clearly, the envelope estimator had significant improvement over the standard method and a slight edge over IRPLS.

5.9 Proofs and technical details

5.9.1 Proposition 5.2 and Corollary 5.1

Proof. The asymptotic normality and the asymptotic covariance of $\widehat{\boldsymbol{\phi}}_n$ can be see by definition. We write the Fisher information $\mathbf{F}(\boldsymbol{\theta}_t)$ as its natural block matrix structure

$$\mathbf{F}(\boldsymbol{\theta}_t) = \mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\phi}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix},$$

where $\mathbf{A} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T\right)\right\}$, $\mathbf{B} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\phi}^T\right)\right\}$ and $\mathbf{C} = -\mathrm{E}_{\boldsymbol{\theta}_t}\left\{\left(\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi})/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T\right)\right\}$. Then $\mathbf{V}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) = \mathbf{F}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t) = (\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}$ by block matrix inversion.

The Hessian matrix for the constrained optimization (5.3.1)-(5.3.3) is

$$\mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\eta}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B}\boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^T\mathbf{B}^T & \boldsymbol{\Gamma}^T\mathbf{C}\boldsymbol{\Gamma} \end{pmatrix}, \tag{5.9.1}$$

since $\partial/\partial\boldsymbol{\eta} = \boldsymbol{\Gamma}^T \cdot \partial/\partial\boldsymbol{\phi}$ under the constraint that $\boldsymbol{\phi} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. Therefore the asymptotic variance for $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is

$$\begin{aligned} \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\boldsymbol{\Gamma}}) &= \boldsymbol{\Gamma}\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{C}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}\left\{\boldsymbol{\Gamma}^T\mathbf{V}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\boldsymbol{\theta}_t)\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}^T \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0\mathbf{V}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\mathbf{V}\mathbf{P}_{\boldsymbol{\Gamma}}, \end{aligned}$$

where the last equality is obtained by applying the equality (Cook and Forzani 2009),

$$(\boldsymbol{\Gamma}^T\mathbf{V}^{-1}\boldsymbol{\Gamma})^{-1} = \boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T\mathbf{V}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\mathbf{V}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\mathbf{V}\boldsymbol{\Gamma}. \tag{5.9.2}$$

Then by requiring that span($\mathbf{\Gamma}$) reduces $\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)$, we have avar$(\sqrt{n}\widehat{\boldsymbol{\phi}}_{\mathbf{\Gamma}}) = \mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}\mathbf{P}_{\mathbf{\Gamma}}$.

Similarly for the asymptotic variance of $\widehat{\boldsymbol{\psi}}_n$ and $\widehat{\boldsymbol{\psi}}_{\mathbf{\Gamma}}$ we have the following results from block matrix inversion and (5.9.2).

$$\begin{aligned}\text{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_{\mathbf{\Gamma}}) &= \left(\mathbf{A} - \mathbf{B}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{C}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{B}^T\right)^{-1} \\ &\leq \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1} = \text{avar}(\sqrt{n}\widehat{\boldsymbol{\psi}}_n).\end{aligned}$$

□

5.9.2 Proposition 5.3

From Cook and Zhang (2014; Proposition 6), we know that $\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}^T$ is $\sqrt{n}$-consistent for the projection onto the envelope $\mathcal{E}_{\mathbf{V}_{\phi\phi}}(\boldsymbol{\phi}_t)$. Then the proof follows from the proof of $\sqrt{n}$-consistency of the k-th direction in the 1D algorithm, where we replace the sample objective function $J_k(\mathbf{g})$ with $L_n(\boldsymbol{\psi}, \widehat{\mathbf{\Gamma}}\boldsymbol{\eta})$.

5.9.3 Proposition 5.4

Proof. The envelope of interest is $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t)$. The Fisher information per observation is then obtained by taking the expectation of the second derivative of $\log f(y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X}) + \log g(x|\boldsymbol{\psi})$ with respect to $\boldsymbol{\theta}$. Clearly, the information matrix is block diagonal in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{\psi}$ so we need only the second derivatives of $\log f(y|\boldsymbol{\alpha}, \boldsymbol{\beta}^T\mathbf{X})$. Then the Fisher information matrix for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be written unambiguously in the form

$$\mathbf{F}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \text{E}\begin{pmatrix} \mathbf{A}(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t) & \mathbf{a}(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}^T \\ \mathbf{X}\mathbf{a}^T(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t) & c(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}\mathbf{X}^T \end{pmatrix},$$

where $\mathbf{A}$ is the second derivative matrix with respect to $\boldsymbol{\alpha}$, the off diagonal blocks are the cross derivatives $\partial^2/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^T$ with all factors collected into $\mathbf{a}$ except $\mathbf{X}$, and $c(Y, \boldsymbol{\beta}_t^T\mathbf{X}, \boldsymbol{\alpha}_t)\mathbf{X}\mathbf{X}^T$ is the second derivative with respect to $\boldsymbol{\beta}$. The asymptotic covariance matrix for $\widehat{\boldsymbol{\beta}}_n$ is then

$$\mathbf{V}_{\beta\beta} = \{\text{E}(c\mathbf{X}\mathbf{X}^T) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\mathbf{X}^T)\}^{-1}.$$

Since $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\mathbf{V}^{-1}_{\beta\beta}}(\boldsymbol{\beta}_t)$, it is sufficient to consider, making the additional assumption that $E(\mathbf{X}|\mathbf{\Gamma}^T\mathbf{X})$ is a linear function of $\mathbf{X}$ and requiring that span$(\boldsymbol{\beta}_t) \subseteq$ span$(\mathbf{\Gamma})$,

$$\begin{aligned}\mathbf{P}_{\mathbf{\Gamma}}\mathbf{V}^{-1}_{\beta\beta}(\boldsymbol{\beta}_t)\mathbf{Q}_{\mathbf{\Gamma}} &= \mathbf{P}_{\mathbf{\Gamma}}\{\text{E}(c\mathbf{X}\mathbf{X}^T) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{Q}_{\mathbf{\Gamma}} \\ &= \mathbf{P}_{\mathbf{\Gamma}}\{\text{E}(c\mathbf{\Gamma}\mathbf{\Gamma}^T\mathbf{X}\cdot\mathbf{X}^T) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{Q}_{\mathbf{\Gamma}} \\ &= \mathbf{P}_{\mathbf{\Gamma}}\{\text{E}(c\mathbf{X}\text{E}(\mathbf{X}^T|Y, \mathbf{\Gamma}^T\mathbf{X})) - \text{E}(\mathbf{X}\mathbf{a}^T)\text{E}^{-1}(\mathbf{A})\text{E}(\mathbf{a}\text{E}(\mathbf{X}^T|Y, \mathbf{\Gamma}^T\mathbf{X}))\}\mathbf{Q}_{\mathbf{\Gamma}}.\end{aligned}$$

This follows since c and $\mathbf{a}$ are functions of only Y and $\boldsymbol{\beta}^T\mathbf{X}$ and span$(\boldsymbol{\beta}_t) \subseteq$ span$(\mathbf{\Gamma})$. Next, using the fact that we are in a regression and thus $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$ implies $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{\Gamma}^T\mathbf{X}$,

we have

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{Q}_{\boldsymbol{\Gamma}} &= \mathbf{P}_{\boldsymbol{\Gamma}}\{\mathrm{E}(c\mathbf{X}\mathrm{E}(\mathbf{X}^T|\boldsymbol{\Gamma}^T\mathbf{X})) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathrm{E}(\mathbf{X}^T|\boldsymbol{\Gamma}^T\mathbf{X}))\}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\{\mathrm{E}(c\mathbf{X}\mathbf{X}^T\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathbf{X}^T\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})})\}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\{\mathrm{E}(c\mathbf{X}\mathbf{X}^T) - \mathrm{E}(\mathbf{X}\mathbf{a}^T)\mathrm{E}^{-1}(\mathbf{A})\mathrm{E}(\mathbf{a}\mathbf{X}^T)\}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_\mathbf{X})}\mathbf{Q}_{\boldsymbol{\Gamma}} \\
&= \boldsymbol{\Gamma}\{\boldsymbol{\Gamma}^T\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\boldsymbol{\Gamma}\}\{\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\boldsymbol{\Gamma}\}^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_{\boldsymbol{\Gamma}}.
\end{aligned}$$

The second equality above uses the linearity condition and the rest is algebra. Since $\boldsymbol{\Gamma}$ has full column rank and the next two matrix factors in $\{\cdot\}$ are positive definite, we see that $\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\mathbf{Q}_{\boldsymbol{\Gamma}} = 0$ if and only if $\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_{\boldsymbol{\Gamma}} = 0$; that is, if and only if span($\boldsymbol{\Gamma}$) reduces $\boldsymbol{\Sigma}_\mathbf{X}$. Consequently, span($\boldsymbol{\Gamma}$) reduces $\mathbf{V}_{\boldsymbol{\beta\beta}}$ if and only if span($\boldsymbol{\Gamma}$) reduces $\boldsymbol{\Sigma}_\mathbf{X}$. Note that in this proof we did not require $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be orthogonal parameters, nor did we require an exponential family regression. The only requirements are the regression structure and the linearity condition. □

5.9.4 Lemma 5.1

Proof. The multivariate normal log-likelihood up to a constant is

$$M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X}) = -\frac{n}{2}\{\log|\boldsymbol{\Sigma}_\mathbf{X}| + \mathrm{trace}[\boldsymbol{\Sigma}_\mathbf{X}^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})^T/n]\}, \tag{5.9.3}$$

Since span($\boldsymbol{\Gamma}$) = $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_t)$ reduces $\boldsymbol{\Sigma}_\mathbf{X}$, we need to maximize $M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ under the constraint that span($\boldsymbol{\Gamma}$) reduces $\boldsymbol{\Sigma}_\mathbf{X}$. This can be done by substitute $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ into the objective function, where $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are two symmetric positive definite matrices. The resulting objective function $M_n(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma})$ can be first maximized over $\boldsymbol{\mu}_\mathbf{X}$ to get $\widehat{\boldsymbol{\mu}}_\mathbf{X} = \overline{\mathbf{X}}$. Then the partially maximized objective function can be expressed as

$$\begin{aligned}
-\frac{2}{n}M_n(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma}) &= \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| + \mathrm{trace}[(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}\mathbf{S}_\mathbf{X}] \\
&= \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \mathrm{trace}\{(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\mathbf{S}_\mathbf{X}\} \\
&= \log|\boldsymbol{\Omega}| + \mathrm{trace}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}) + \log|\boldsymbol{\Omega}_0| + \mathrm{trace}\{(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}) \\
&= \log|\boldsymbol{\Omega}| + \mathrm{trace}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}) + \log|\boldsymbol{\Omega}_0| + \mathrm{trace}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0).
\end{aligned}$$

For any symmetric positive definite matrix $\mathbf{M}$, we apply the following result to optimize over $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ in $M_n(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma})$,

$$\arg\min_{\mathbf{A}}\{\log|\mathbf{A}| + \mathrm{trace}(\mathbf{A}^{-1}\mathbf{M})\} = \mathbf{M}, \tag{5.9.4}$$

where the minimization is over all positive definite symmetric matrices. Therefore we have $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}$, $\widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0$ and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}}\boldsymbol{\Gamma}_0^T = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{S}_\mathbf{X}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{S}_\mathbf{X}\mathbf{Q}_{\boldsymbol{\Gamma}}. \tag{5.9.5}$$

The remaining partially maximized form is $M_n(\boldsymbol{\Gamma}) = -n/2\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0|\right\}$ where we have omitted a constant p from $\text{trace}\{\mathbf{I}_u\} + \text{trace}\{\mathbf{I}_{p-u}\}$. The rest of the proof is a direct consequence of the following implication from Cook et al. (2013): Suppose $\mathbf{A} \in \mathbb{S}^{t\times t}$ is non-singular and that the column partition $(\mathbf{B}, \mathbf{B}_0) \in \mathbb{R}^{t\times t}$ is orthogonal. Then (1) $|\mathbf{B}^T\mathbf{A}\mathbf{B}| = |\mathbf{A}| \cdot |\mathbf{B}_0^T\mathbf{A}^{-1}\mathbf{B}_0|$ and (2) $|\mathbf{A}| \leq |\mathbf{B}^T\mathbf{A}\mathbf{B}| \cdot |\mathbf{B}_0^T\mathbf{A}\mathbf{B}_0|$ if and only if $\text{span}(\mathbf{B})$ reduces $\mathbf{A}$. □

5.9.5 Proposition 5.5

Proof. Let $\mathbf{E}_p \in \mathbb{R}^{p^2\times p(p+1)/2}$ denote the expansion operator and let $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2\times p^2}$ denote the contraction operator such that they connect vec and vech operators as $\text{vec}(\mathbf{A}) = \mathbf{E}_p\text{vech}(\mathbf{A})$ and $\text{vech}(\mathbf{A}) = \mathbf{C}_p\text{vec}(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{S}^{p\times p}$.

Because the standard estimator $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{\Sigma}}_\mathbf{X} = \mathbf{S}_\mathbf{X}$ were obtained separately from $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the marginal multivariate normal likelihood of $\mathbf{X}$, and because that the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal, the asymptotic variance of $\widehat{\boldsymbol{h}}_n$ is block diagonal as $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\alpha}}_n) = \mathbf{V}(\boldsymbol{\alpha}_t) = \mathbf{F}^{-1}(\boldsymbol{\alpha}_t)$, $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_n) = \mathbf{V}(\boldsymbol{\beta}_t) = \mathbf{F}^{-1}(\boldsymbol{\beta}_t)$ and $\text{avar}[\sqrt{n}\text{vech}(\widehat{\boldsymbol{\Sigma}}_\mathbf{X})] = \mathbf{F}_{\boldsymbol{\Sigma}_\mathbf{X}}^{-1}$, where the Fisher information matrix $\mathbf{F}_{\boldsymbol{\Sigma}_\mathbf{X}} = \mathbf{E}_p^T(\boldsymbol{\Sigma}_\mathbf{X}^{-1} \otimes \boldsymbol{\Sigma}_\mathbf{X}^{-1})\mathbf{E}_p/2$. (Cook et al. 2013). Hence,

$$\begin{aligned}\mathbf{F}(\boldsymbol{h}_t) &= \text{E}\left\{\left(\frac{\partial L_n(\boldsymbol{h})}{\partial \boldsymbol{h}}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\left(\frac{\partial L_n(\boldsymbol{h})}{\partial \boldsymbol{h}^T}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\right\} = -\text{E}\left\{\left(\frac{\partial^2 L_n(\boldsymbol{h})}{\partial \boldsymbol{h}\partial \boldsymbol{h}^T}\right)_{\boldsymbol{h}=\boldsymbol{h}_t}\right\} \\ &= \text{diag}\left(\mathbf{F}(\boldsymbol{\alpha}_t), \mathbf{F}(\boldsymbol{\beta}_t), \mathbf{F}_{\boldsymbol{\Sigma}_\mathbf{X}}\right).\end{aligned}$$

Following Cook et al. (2013), the asymptotic variance for $\widehat{\boldsymbol{h}}_{\text{env}} = \boldsymbol{h}(\widehat{\boldsymbol{\phi}}_{\text{env}})$ can be expressed as $\text{avar}(\widehat{\boldsymbol{h}}_{\text{env}}) = \mathbf{H}(\mathbf{H}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H})^\dagger\mathbf{H}^T$, where $\mathbf{H} = \partial\boldsymbol{h}/\partial\boldsymbol{\phi} = (\partial\boldsymbol{h}_i/\partial\boldsymbol{\phi}_j)_{i=1,\ldots,3,j=1,\ldots,5}$ is the gradient matrix and † denotes the Moore-Penrose generalized inverse. By direct computation, we found

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \tag{5.9.6}$$

Since $\text{avar}(\widehat{\boldsymbol{h}}_{\text{env}}) = \mathbf{H}(\mathbf{H}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H})^\dagger\mathbf{H}^T$ depends only on $\text{span}(\mathbf{H})$, we transform $\mathbf{H} \to \mathbf{H}_1$ similar to Cook et al. (2013) so that $\text{span}(\mathbf{H}) = \text{span}(\mathbf{H}_1)$ and

$$\text{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{env}}) = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_1)^\dagger\mathbf{H}_1^T = \sum_{j=1}^{5}\mathbf{H}_{1j}(\mathbf{H}_{1j}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{1j})^\dagger\mathbf{H}_{1j}^T, \tag{5.9.7}$$

where $\mathbf{H}_1 = (\mathbf{H}_{11}, \ldots, \mathbf{H}_{15})$ has the blockwise form of

$$\begin{aligned}\mathbf{H}_1 &= \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} & \mathbf{H}_{15} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}.\end{aligned}$$

Because of the zeros blocks in $\mathbf{H}_{11}$, $\mathbf{H}_{14}$ and $\mathbf{H}_{15}$, they have no contribution to the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{env}}$, which is the middle $p \times p$ block of avar$(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{env}})$:

$$\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) = \boldsymbol{\Gamma}(\mathbf{H}_{12}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{12})^\dagger\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)(\mathbf{H}_{13}^T\mathbf{F}(\boldsymbol{h}_t)\mathbf{H}_{13})^\dagger(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T). \quad (5.9.8)$$

Then by straightforward calculations similar to Cook et al. (2010), we have

$$\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) = \boldsymbol{\Gamma}\left\{\boldsymbol{\Gamma}^T\mathbf{F}(\boldsymbol{\beta}_t)\boldsymbol{\Gamma}\right\}^{-1}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T),$$

where $\mathbf{F}(\boldsymbol{\beta}_t) = \left\{\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_n)\right\}^{-1}$ and $\mathbf{M} = (\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^T)\mathbf{F}(\boldsymbol{\beta}_t)(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_{u(p-u)}$.

From Proposition 5.2, we recognize the first term in avar$(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}})$ is avar$(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ which equals to avar$(\sqrt{n}\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})$. To further interpret the second term, we follow the same calculation as in Cook et al. (2010) by making the following substitutes: $\boldsymbol{\Sigma} \to \boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} \to \mathbf{F}(\boldsymbol{\beta}_t)$ and the dimension $r \to p$. Then the conclusion follows.

□

5.9.6 Corollary 5.2

Proof. From Proposition 5.4, we can see $\mathcal{E}_{\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\sigma^2\mathbf{I}_p}(\boldsymbol{\beta}_t) = \text{span}(\boldsymbol{\beta}_t)$. Therefore, the envelope is one-dimensional and $\boldsymbol{\beta}_t = \boldsymbol{\Gamma}\eta$ with $\boldsymbol{\Gamma} = \boldsymbol{\beta}_t/||\boldsymbol{\beta}_t|| \in \mathbb{R}^{p\times 1}$ and $\eta = ||\boldsymbol{\beta}_t|| \in \mathbb{R}^1$. The covariances $\Omega = \sigma^2$ and $\boldsymbol{\Omega}_0 = \sigma^2\mathbf{I}_{p-1}$. The matrix $\mathbf{M}$ in Proposition 5.5 becomes $\mathbf{M} = (\eta\boldsymbol{\Gamma}_0^T)\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\eta\boldsymbol{\Gamma}_0)$ because $\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 = \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} = \mathbf{I}_{p-1}$ cancel with the last term. Hence,

$$\begin{aligned}\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{env}}) &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + (\eta\boldsymbol{\Gamma}_0)\mathbf{M}^{-1}(\eta\boldsymbol{\Gamma}_0^T) \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + \eta^2\boldsymbol{\Gamma}_0(\eta^2\boldsymbol{\Gamma}_0^T\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T \\ &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{Q}_{\boldsymbol{\Gamma}} \\ &= \mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}),\end{aligned}$$

where the last two equalities have used the fact that $\boldsymbol{\Gamma}$ reduces $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}$. □

5.9.7 Proposition 5.1

Proof. Let $(\mathbf{R} \otimes \mathbf{L})_0 \equiv (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})$ is a $pr \times (pr - d_X d_Y)$ semi-orthogonal matrix so that $\mathbf{O} \equiv (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L}, \mathbf{R} \otimes \mathbf{L})$ is an orthogonal basis for $\mathbb{R}^{pr}$, where $(\mathbf{R}, \mathbf{R}_0)$ and $(\mathbf{L}, \mathbf{L}_0)$ are orthogonal bases for $\mathbb{R}^p$ and $\mathbb{R}^r$.

We first prove that the following statements are equivalent.

1. span($\mathbf{R} \otimes \mathbf{L}$) reduce $\mathbf{\Delta}_R \otimes \mathbf{\Delta}_L$.
2. span($\mathbf{R}$) reduces $\mathbf{\Delta}_R$ and span($\mathbf{L}$) reduces $\mathbf{\Delta}_L$.
3. $\mathbf{O}(\mathbf{\Delta}_R \otimes \mathbf{\Delta}_L)\mathbf{O}^T = \text{diag}\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, where $\mathbf{M}_1 = \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}$, $\mathbf{M}_2 = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0^T \otimes \mathbf{L}_0\mathbf{\Delta}_L\mathbf{L}_0^T$, $\mathbf{M}_3 = \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}^T \otimes \mathbf{L}_0\mathbf{\Delta}_L\mathbf{L}_0^T$ and $\mathbf{M}_4 = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0^T \otimes \mathbf{L}\mathbf{\Delta}_L\mathbf{L}^T$.
4. $\mathbf{O}(\mathbf{\Delta}_R \otimes \mathbf{\Delta}_L)\mathbf{O}^T = \text{diag}\{\mathbf{M}_1, \mathbf{M}_0\}$, where $\mathbf{M}_0 = (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})(\mathbf{\Delta}_R \otimes \mathbf{\Delta}_L)(\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})^T$.

By definition, we know that Statement 1 is equivalent to 4 and that 2 is equivalent to 3. Also, 3 implies 4. We then only need to show that 4 implies 3. From 4, we know that the off-diagonal blocks of $\mathbf{O}(\mathbf{\Delta}_R \otimes \mathbf{\Delta}_L)\mathbf{O}^T$ are all zero matrices. More explicitly,

$$\begin{aligned} 0 &= \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L} \\ &= \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L} = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L} \\ &= \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0. \end{aligned} \tag{5.9.9}$$

Without loss of generality, we assume that $\mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}$, $\mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}$, $\mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0$ and $\mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L}_0$ are all nonzero. Otherwise, the problem degenerates to the case of reducing either $\mathbf{\Delta}_R$ or $\mathbf{\Delta}_L$ and the proof would be trivial. Because $\mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}$, $\mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}$, $\mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0$ and $\mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L}_0$ are all nonzero, we must have from (5.9.9) that

$$0 = \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L} = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R} = \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}_0,$$

which implies the following blocks within $\mathbf{M}_0$ are all zero matrices:

$$\begin{aligned} 0 &= \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L} \\ &= \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}_0^T\mathbf{\Delta}_L\mathbf{L}_0 \\ &= \mathbf{R}_0^T\mathbf{\Delta}_R\mathbf{R} \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0 = \mathbf{R}^T\mathbf{\Delta}_R\mathbf{R}_0 \otimes \mathbf{L}^T\mathbf{\Delta}_L\mathbf{L}_0. \end{aligned}$$

Hence, $\mathbf{M}_0 = \text{diag}\{\mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, which leads to the equivalence of the four statements. From Definition 5.1, Definition 5.2 and the equivalence between statement 1 and 2, we reach the conclusion of Proposition 5.1.

□

5.9.8 Derivations for equation (5.5.6)

The normality assumption $Y|(\mathbf{X}, W) \sim N(\mu + \boldsymbol{\beta}^T\mathbf{X}, \sigma^2/W)$ leads to the following conditional log-likelihood:

$$C_n(\mu, \boldsymbol{\beta}, \sigma^2) = -n/2\log(\sigma^2) - \sum_{i=1}^{n} W_i(Y_i - \mu - \boldsymbol{\beta}^T\mathbf{X}_i)/(2\sigma^2). \tag{5.9.10}$$

By straightforward computation, the maximum likelihood estimators is obtained as:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \mathbf{S}_{\mathbf{X}(W)}^{-1}\mathbf{S}_{\mathbf{X}Y(W)} && (5.9.11)\\
\widehat{\mu} &= \sum_{i=1}^{n} W_i Y_i/n - \widehat{\boldsymbol{\beta}}^T \sum_{i=1}^{n} W_i \mathbf{X}_i/n \equiv \overline{Y}_{(W)} - \widehat{\boldsymbol{\beta}}^T \overline{\mathbf{X}}_{(W)} && (5.9.12)\\
\widehat{\sigma}^2 &= s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \mathbf{S}_{\mathbf{X}Y(W)} \equiv s^2 && (5.9.13)
\end{aligned}$$

Given $\boldsymbol{\Gamma}$, which is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t)$, we then only need to maximize $C_n(\mu, \boldsymbol{\beta}, \sigma^2)$ under the constraint that $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$. The resulting estimators are:

$$\begin{aligned}
\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} &= (\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}Y(W)} && (5.9.14)\\
\widehat{\mu}_{\boldsymbol{\Gamma}} &= \overline{Y}_{(W)} - \widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}^T \boldsymbol{\Gamma}^T \overline{\mathbf{X}}_{(W)} && (5.9.15)\\
\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}} = \mathbf{P}_{\boldsymbol{\Gamma}(\mathbf{S}_{\mathbf{X}(W)})}\widehat{\boldsymbol{\beta}} && (5.9.16)\\
\widehat{\sigma}_{\boldsymbol{\Gamma}}^2 &= s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}Y(W)} \equiv s_{\boldsymbol{\Gamma}}^2. && (5.9.17)
\end{aligned}$$

If we replace $\boldsymbol{\Gamma}$ by its $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\Gamma}}$ from either a moment-based objective function or a likelihood-based objective function, then the above estimators will be our envelope estimators.

To obtain a likelihood-based objective function for $\boldsymbol{\Gamma}$, we need to compute the partially maximized conditional likelihood of $Y|(\mathbf{X}, W)$: $C_n(\boldsymbol{\Gamma}) = C_n(\widehat{\mu}_{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\Gamma}})$, plus the partially maximized marginal likelihood $M_n(\boldsymbol{\Gamma})$ of $\mathbf{X}|W$. First, we can get $C_n(\boldsymbol{\Gamma})$ from above derivations as:

$$C_n(\boldsymbol{\Gamma}) = -n/2 \log(s_{\boldsymbol{\Gamma}}^2) - n/2. \qquad (5.9.18)$$

Assuming that $\mathbf{X}$ and W are independent and that $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, we have by Proposition 5.4 that $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t)$. The partially maximized log-likelihood $M_n(\boldsymbol{\Gamma})$ follows from Lemma 5.1:

$$M_n(\boldsymbol{\Gamma}) = -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}. \qquad (5.9.19)$$

Therefore, the final objective function for $\boldsymbol{\Gamma}$ is

$$\begin{aligned}
L_n(\boldsymbol{\Gamma}) &= C_n(\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma})\\
&= -\frac{n}{2}\left\{1 + \log(s_{\boldsymbol{\Gamma}}^2) + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}}|\right\}\\
&= -\frac{n}{2}\log\left\{s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}Y(W)}\right\}\\
&\quad -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}|\right\} + \text{constant}. && (5.9.20)
\end{aligned}$$

Alternatively, if we assume $\mathbf{X}|W \sim N(\boldsymbol{\mu}_{\mathbf{X}}, W^{-1}\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$ and use the fact that

$\mathrm{E}(W) = 1$, then $\boldsymbol{\mu}_{\mathbf{X}} = \mathrm{E}(W\mathbf{X}) = \mathrm{E}(\mathbf{X})$. Moreover, the covariance

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{X}(W)} &= \mathrm{E}\left[W(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^T\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[W(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})|W\right]\right\} \\ &= \mathrm{E}\left\{W\cdot\mathrm{var}(\mathbf{X}|W)\right\} = \mathrm{E}(W\cdot W^{-1}\boldsymbol{\Sigma}) \\ &= \boldsymbol{\Sigma}.\end{aligned}$$

The maximum likelihood estimators for $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}$ are then $\widehat{\boldsymbol{\mu}}_{\mathbf{X}} = \overline{\mathbf{X}}_{(W)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}(W)}$. Follow the same calculation in the proof of Lemma 5.1, we have

$$M_n(\boldsymbol{\Gamma}) = -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{X}(W)}|\right\}.$$

Therefore, the final objective function for $\boldsymbol{\Gamma}$ up to a constant is

$$\begin{aligned}L_n(\boldsymbol{\Gamma}) &= C_n(\boldsymbol{\Gamma}) + M_n(\boldsymbol{\Gamma}) \\ &= -\frac{n}{2}\left\{\log(s^2_{\boldsymbol{\Gamma}}) + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}|\right\} \\ &= -\frac{n}{2}\log\left\{s^2_{Y(W)} - \mathbf{S}^T_{\mathbf{X}Y(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}Y(W)}\right\} \\ &\quad -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}|\right\} \\ &= -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\left(\mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)}s^{-2}_{Y(W)}\mathbf{S}^T_{\mathbf{X}Y(W)}\right)\boldsymbol{\Gamma}|\right\} \\ &\equiv -\frac{n}{2}\left\{\log|\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}(W)}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}|Y(W)}\boldsymbol{\Gamma}|\right\}.\end{aligned} \tag{5.9.21}$$

5.9.9 Implementation details for envelope GLM in Section 5.5.2

Let $\mathbb{X}_n$ be the $n\times p$ data matrix of $\mathbf{X}_i$'s and let $\mathbb{C}_n$ be the $n\times 1$ data matrix of $\mathcal{C}'(\vartheta_i)$, where $\mathcal{C}'(\vartheta)$ is the first order derivative from Table 5.1. Then the $p\times u$ gradient matrix of $\partial C_n(\boldsymbol{\Gamma})/\partial\boldsymbol{\Gamma}$ can be computed conveniently as follows.

Lemma 5.2.

$$\begin{aligned}\frac{\partial C_n(\boldsymbol{\Gamma})}{\partial\boldsymbol{\Gamma}} &= \mathbb{X}_n^T\mathbb{C}_n\boldsymbol{\eta}^T + \mathbf{S}_{\mathbf{X}V(W)}\mathbb{C}_n^T\mathbb{X}_n\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1} \\ &\quad -\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbb{X}_n^T\mathbb{C}_n\mathbf{S}^T_{\mathbf{X}V(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1} \\ &\quad -\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)}\mathbb{C}_n^T\mathbb{X}_n\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\boldsymbol{\Gamma})^{-1}; \\ \frac{\partial M_n(\boldsymbol{\Gamma})}{\partial\boldsymbol{\Gamma}} &= -n\left\{\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1} + \mathbf{S}^{-1}_{\mathbf{X}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}^{-1}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}\right\}.\end{aligned}$$

To check this Lemma, we compared the analytical form of the derivatives to the numerical derivatives. The numerical derivatives of a function $f(\boldsymbol{\Gamma})$ is computed by varying each element of $\boldsymbol{\Gamma}$ by a small number $\delta = 1.0\times 10^{-6}$ and evaluate the function:

$$\left[\frac{\partial f(\boldsymbol{\Gamma})}{\partial\boldsymbol{\Gamma}}\right]_{ij} = \frac{f(\boldsymbol{\Gamma}+\delta\mathbf{G}_{ij}) - f(\boldsymbol{\Gamma}-\delta\mathbf{G}_{ij})}{2\delta}, \tag{5.9.22}$$

where $\mathbf{G}_{ij} \in \mathbb{R}^{p\times u}$ is zero matrix except its ij-th entry is one. We found the difference between our analytical derivative and the numerical derivative is around 10^{-7} for every element.

Proof. This proof of Lemma 5.2 mainly involves rather complicated matrix differentiation.

$$\begin{aligned}
\frac{\partial \vartheta_i}{\partial \mathrm{vec}(\mathbf{\Gamma})} &= \frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})}\left\{\mathbf{X}_i^T\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)}\right\} \\
&= \frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})}\left\{\left(\mathbf{S}_{\mathbf{X}V(W)}^T\otimes\mathbf{X}_i^T\right)\mathrm{vec}[\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T]\right\} \\
&= \left[\frac{\partial}{\partial \mathrm{vec}(\mathbf{\Gamma})}\mathrm{vec}(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)\right]^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right). \qquad (5.9.23)
\end{aligned}$$

Apply the following chain rule:

$$\partial\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T\mathbf{B}^T\otimes\mathbf{I})\partial\mathrm{vec}(\mathbf{A}) + (\mathbf{C}^T\otimes\mathbf{A})\partial\mathrm{vec}(\mathbf{B}) + (\mathbf{I}\otimes\mathbf{AB})\partial\mathrm{vec}(\mathbf{C}),$$

we get

$$\begin{aligned}
\partial\mathrm{vec}(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)/\partial\mathrm{vec}(\mathbf{\Gamma}) &= \mathbf{T}_1+\mathbf{T}_2+\mathbf{T}_3 \\
&= \left[\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\otimes\mathbf{I}_p\right]\mathbf{I}_{pu} \\
&\quad +(\mathbf{\Gamma}\otimes\mathbf{\Gamma})\frac{\partial\mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}}{\partial\mathrm{vec}(\mathbf{\Gamma})} \\
&\quad +\left[\mathbf{I}_p\otimes\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right]\mathbf{K}_{pu}.
\end{aligned}$$

The second term after the last equal sign is computed by noticing that

$$\begin{aligned}
\partial\mathrm{vec}\left\{(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right\} &= -\mathrm{vec}\left\{(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\left[\partial(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})\right](\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right\} \\
&= -\left((\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\otimes(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right)\partial\mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}),
\end{aligned}$$

and that

$$\begin{aligned}
\frac{\partial\mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})}{\partial\mathrm{vec}(\mathbf{\Gamma})} &= (\mathbf{I}_u\otimes\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)})\frac{\partial\mathrm{vec}(\mathbf{\Gamma})}{\partial\mathrm{vec}(\mathbf{\Gamma})}+(\mathbf{\Gamma}^T\otimes\mathbf{I}_u)\frac{\partial\mathrm{vec}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)})}{\partial\mathrm{vec}(\mathbf{\Gamma})} \\
&= (\mathbf{I}_u\otimes\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)})+(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\otimes\mathbf{I}_u)\mathbf{K}_{pu}.
\end{aligned}$$

Hence the second term equals to

$$\begin{aligned}
\mathbf{T}_2 &= -\left(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\otimes\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right) \\
&\quad -\left(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\otimes\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right)\mathbf{K}_{pu}.
\end{aligned}$$

Finally, substitute the three terms into (5.9.23), we have the three terms from

$$\begin{aligned}
\frac{\partial\vartheta_i}{\partial\mathrm{vec}(\mathbf{\Gamma})} &= (\mathbf{T}_1+\mathbf{T}_2+\mathbf{T}_3)^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right) \\
\mathbf{T}_1^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right) &= \left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right] \\
&= [\boldsymbol{\eta}\otimes\mathbf{X}_i] \\
\mathbf{T}_3^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right) &= \left\{\left[\mathbf{I}_p\otimes\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\right]\mathbf{K}_{pu}\right\}^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right) \\
&= \mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}V(W)}\otimes(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right]
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}_2^T\left(\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{X}_i\right) &= -\left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)}\otimes\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\
&\quad -\mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}V(W)}\otimes(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\
&= -\left[\boldsymbol{\eta}\otimes\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right] \\
&\quad -\mathbf{K}_{pu}^T\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}\boldsymbol{\eta}\otimes(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\right]
\end{aligned}$$

Notice that each of the above terms is a $pu\times 1$ vector, and then the $\mathbf{K}_{pu}^T$ serves only to change $\mathrm{vec}(\mathbf{A})\to\mathrm{vec}(\mathbf{A}^T)$, therefore

$$\begin{aligned}
\frac{\partial\vartheta_i}{\partial\mathrm{vec}(\mathbf{\Gamma})} &= \mathrm{vec}(\mathbf{M}_1+\mathbf{M}_3^T)+\mathrm{vec}(\mathbf{M}_2+\mathbf{M}_4^T) \\
\mathbf{M}_1 &= \left[\mathbf{X}_i\cdot\boldsymbol{\eta}^T\right] \\
\mathbf{M}_3 &= \left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\cdot\mathbf{S}_{\mathbf{X}V(W)}^T\right] \\
\mathbf{M}_2 &= -\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\cdot\boldsymbol{\eta}^T\right] \\
\mathbf{M}_4 &= -\left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\cdot\boldsymbol{\eta}^T\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right],
\end{aligned}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3^T$, $\mathbf{M}_4^T$ are all $p\times u$ matrices.

To implement, we have the gradient matrix as

$$\frac{\partial C_n(\mathbf{\Gamma})}{\partial\mathbf{\Gamma}}+\frac{\partial M_n(\mathbf{\Gamma})}{\partial\mathbf{\Gamma}}=\frac{\partial C_n(\mathbf{\Gamma})}{\partial\mathbf{\Gamma}}-n\{\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma})^{-1}+\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma})^{-1}\},\qquad(5.9.24)$$

where $\frac{\partial C_n(\mathbf{\Gamma})}{\partial\mathbf{\Gamma}}$ is re-arranged as a $p\times u$ matrix

$$\begin{aligned}
\frac{\partial C_n(\mathbf{\Gamma})}{\partial\mathbf{\Gamma}} &= \left\{\sum_{i=1}^n\mathcal{C}'(\vartheta_i)\frac{\partial\vartheta_i}{\partial\mathrm{vec}(\mathbf{\Gamma})}\right\}_{p\times u} \\
&= \sum_{i=1}^n\mathcal{C}'(\vartheta_i)\{\mathbf{X}_i\cdot\boldsymbol{\eta}^T+\mathbf{S}_{\mathbf{X}V(W)}\cdot\mathbf{X}_i^T\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\} \\
&\quad -\sum_{i=1}^n\mathcal{C}'(\vartheta_i)\left[\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\cdot\boldsymbol{\eta}^T\right] \\
&\quad -\sum_{i=1}^n\mathcal{C}'(\vartheta_i)\left[(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{X}_i\cdot\boldsymbol{\eta}^T\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\right]^T \\
&= \mathbb{X}_n^T\mathbb{C}_n\boldsymbol{\eta}^T+\mathbf{S}_{\mathbf{X}V(W)}\mathbb{C}_n^T\mathbb{X}_n\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1} \\
&\quad -\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbb{X}_n^T\mathbb{C}_n\boldsymbol{\eta}^T \\
&\quad -\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma}\boldsymbol{\eta}\mathbb{C}_n^T\mathbb{X}_n\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}(W)}\mathbf{\Gamma})^{-1}.
\end{aligned}$$

□

Chapter 6

Conclusions and future directions

The objective of an envelope method is to reduce estimative variation in multivariate analysis. The reduction is not necessarily associated with predictors or responses, but is generally defined as effective dimensionality reduction in the parameter space. This reduction is achieved by enveloping the variation in estimation that is material to the goals of the analysis while simultaneously excluding the immaterial variation. Efficiency gains are then achieved by identifying and enveloping the material information in the data. The gains can be massive, sometimes equivalent to taking thousands of additional observations.

Previous work on envelopes has successfully improved efficiency in estimating the coefficient matrix in a multivariate linear model. In this thesis, we deepened and broadened our understanding regarding envelopes. We have addressed the following questions in this thesis.

(1) Given an estimation procedure, how to improve its efficiency using the idea of envelopes?

(2) How to construct an envelope estimator in general?

(3) Estimation of envelopes usually involves optimizations over Grassmann manifolds, which are often challenging. How to deal with multiple local minima and how to find initial values for optimization algorithms were important open issues.

Specifically, we proposed a constructive definition (Chapter 5), a fast and stable algorithm (Chapter 2) and a general framework (Chapter 5) for adapting envelopes to any multivariate estimation procedure. Within the context of multivariate linear models, we applied the idea of envelopes to synchronized predictor and response reduction (Chapter 3) and to reduced-rank regression (Chapter 4). Throughout our exposition, we discovered significant improvements by envelope methods over the standard methods in linear regression, generalized linear models, linear discriminant analysis and Cox regression.

Envelope models and methods are growing to a new area of research. We have studied and explored their fundamental properties in this thesis that opened the doors to many interesting applications and theories. We point out some of these in the following.

Envelope regression is very promising for reducing massive data sets in high-dimensional settings, which have became more and more common in the applied sciences. There are many challenging and important problems awaiting to be solved. We want to know the theoretical properties of the envelope estimators in high-dimensional data, under various assumptions such as the sparsity assumption. Also, is it possible to achieve consistent predictive performance even when the envelope basis is not consistent in high-dimensional settings where $p/n \nrightarrow 0$?

In the previous studies of envelope models and methods, little attention has been paid to incorporating prior information in the problem. However, the estimation of interesting parameters is often constrained based on prior knowledge. The estimated model and interpretation should be consistent with prior information such as group structures in the variables. I am also interested in developing suitable methods that utilize prior information and satisfy theory- or data-driven constraints.

Future applications of envelope methods ranging from biometrics and chemometrics to computer vision and data-mining. Different data types requires different modeling, for example, it is interesting to see how envelopes may applied to longitudinal data, genomic data and tensor-valued data.

All the methods in this thesis are implemented using Matlab and relevant computing codes can be found at `http://users.stat.umn.edu/~zhan0648/computing.html`.

References

[1] Absil, P.A., Mahony, R. and Sepulchrer, R. (2008), Optimization Algorithms on Matrix Manifolds. *Princeton University Press.*

[2] Adragni, K., Cook, R.D. and Wu, S. (2012), GrassmannOptim: An R Package for Grassmann Manifold Optimization, *Journal of Statistical Software*, **50**, 1–18.

[3] Agresti, A. (2002), Categorical data analysis, 2nd Edition. *New York: Wiley.*

[4] Amemiya, T. (1985), Advanced Econometrics, *Harvard University Press.*

[5] Anderson, T.W. (1951), Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, **22**, 327–351.

[6] Anderson, T.W. (1999), Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, **27**, 1141–1154.

[7] Anderson, T.W. (2002), Canonical correlation analysis and reduced rank regression in autoregressive models. *The Annals of Statistics*, **30**, 1134–1154.

[8] Bach, F.R. and Jordan, M.I. (2005), A probabilistic interpretation of canonical correlation analysis, *Technical Report*, **688**, 1–11.

[9] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.

[10] Brown, P.J., Fearny, T. and Vannucciz, M. (2001), Bayesian Wavelet Regression on Curves with Application to a Spectroscopic Calibration Problem, *Journal of the American Statistical Association*, **96**, 398–408.

[11] Bura, E. and Cook, R.D. (2003), Rank estimation in reduced-rank regression, *Journal of Multivariate Analysis*, **87**, 159–176.

[12] Carroll, R.J. and Cline, D.B.H. (1988). An Asymptotic theory for weighted least squares with weights estimated by replication. *Biometrika*, **75**, 35–43.

[13] Chun, H. and Keles, S. (2010), Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B*, **72**, 3–25.

[14] Chen, L. and Huang, J.Z. (2012), Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, *Journal of the American Statistical Association*, **107**, 1533–1545.

[15] Chen, K., Chan, K.S. and Stenseth, N.C. (2012), Reduced rank stochastic regression with a sparse singular value decomposition. *JRSS-B*, **74**, 203–221.

[16] Conway, J. (1990). A Course in Functional Analysis. Second edition. *Springer, New York.*

[17] Cook, R.D. (1996), Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

[18] Cook, R.D. (1998), Regression Graphics: Ideas for Studying Regressions Through Graphics. *New York: Wiley.*

[19] Cook, R.D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.

[20] Cook, R.D., Helland, I.S. and Su, Z. (2013), Envelopes and partial least squares regression. *JRSS-B*, **75**,851–877.

[21] Cook, R.D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, **20**,927–1010.

[22] Cook, R.D. and Weisberg, S. (1994), An Introduction to Regression Graphics. *New York: Wiley.*

[23] Cook, R.D. and Zhang, X. (2014). Simultaneous envelopes for multivariate linear regression. *Technometrics*, DOI:10.1080/00401706.2013.872700.

[24] Cox, D.R. (1972), Regression models and life-tables (with Discussi1987on), *J. R. Statist. Soc. B*, **34**, 187–220.

[25] Cox, D.R. (1975), Partial likelihood, *Biometrika*, **62**, 269–276.

[26] Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society B*, **49**, 1–39.

[27] de Jong, S. (1993), SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*,**18**, 251–26.

[28] Dawid, A.P. (1981), Some matrix-variate distribution theory: Notational considerations and a Bayesian application, *Biometrika*, **68**, 265–274.

[29] Eaton, M.L. (1986), A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.

[30] Edelman, A., Tomas, A.A. and Smith, S.T. (1998), The geometry of algorithms with orthogonality constraints, *SIAM Journal of Matrix Analysis and Applications*, **20**, 303–353.

[31] Efron, B. (1974), The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association*, **72**, 557–565.

[32] Hawkins, D.M. and Maboudou-Tchao, E.M. (2013), Smoothed Linear Modeling for Smooth Spectral Data. *To appear in International Journal of Spectroscopy.*

[33] Henderson, H.V. and Searle, S.R. (1979), Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian Journal of Statistics*, **7**, 65–81.

[34] Hooper, J. (1959), Simultaneous equations and canonical correlation theory. *Econometrica*, **27**, 245–256.

[35] Hotelling, H. (1936), Relations between two sets of variates. *Biometrika.* **28**, 321–377.

[36] Hung, H., Wu, P. S., Tu, I. P. and Hung, S. Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, **99**, 569–583.

[37] Izenman, A. J. (1975), Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis.* **5**, 248–264.

[38] Johnstone, I. and A. Lu (2009), On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**, 682–693.

[39] Jolliffe, I. (2005), Principal Component Analysis. *Encyclopedia of Statistics in Behavioral Science.*

[40] Jolliffe, I. (1986), Principal component analysis, *Springer Verlag, New York.*

[41] Kenward, M. G. (1987), A method for comparing profiles of repeated measurements. *JRSS-C*, **36**, 296–308.

[42] Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation. Second edition. *Springer, New York.*

[43] Li, B., Kim, M.K. and Altman, M. (2010). On dimension folding of matrix or array valued statistical objects. *Annals of Statistics*, **38**, 1097–1121.

[44] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association*, **102**, 997–1008.

[45] Li, B., Wen, S. and Zhu, L.X. (2008) On a Projective Resampling Method for Dimension Reduction With Multivariate Responses, *Journal of the American Statistical Association*, **103**, 1177–1186.

[46] Li, K.C., Aragon, Y., Shedden, K. and Agan, C.T. (2003), Dimension reduction for multivariate response data, *Journal of the American Statistical Association*, **98**, 99–109.

[47] Li, K. C. and Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics*,**17**, 1009–1052.

[48] Marx, B. D. (1996), Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*,**38**, 374–381.

[49] Nygard, S. and Borgan, O. (2008), Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*,**14**, 179–195.

[50] Reinsel, G. C. and Velu, R. P. (1998), Multivariate reduced-rank regression: theory and applications. *Springer, New York.*

[51] Rencher, A.C. (2002), Methods of Multivariate Analysis, 2nd Ed. New York, NY: Wiley.

[52] Schomacker, K.T., Frisoli, J.K., Compton, C.C., Flotte, T.J., Richter, J.M., Nishioka, N.S. and Deutsch, T.F. (1992), Ultraviolet laser-induced fluorescence of colonic tissue: basic biology and diagnostic potential. *Lasers in Surgery and Medicine*, **12**, 63–78.

[53] Schott, J.R. (2013), On the likelihood ratio test for envelope models in multivariate linear regression. *Biometrika*, **100**, 531–537.

[54] Seber, G.A.F. (2008), A matrix handbook for statisticians, *Wiley-Interscience*.

[55] Shao, J. (1997), An asymptotic theory for linear model selection (with discussion). *Statistica sinica*, **7**, 221–264.

[56] Shapiro, A. (1986), Asymptotic theory of overparameterized structural models, *Journal of the American Statistical Association*, **81**, 142–149.

[57] Small, C.G., Wang, J. and Yang, Z. (2000) Eliminating multiple root problems in estimation. *Statistical Science*, **15**, 313–341.

[58] Stoica, P. and Viberg, M. (1996), Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions. *IEEE Transactions on Signal Processing*, **44**, 3069–3079.

[59] Su, Z. and Cook, R.D. (2011), Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, **98**, 133–146.

[60] Su, Z. and Cook, R.D. (2012), Inner envelopes: efficient estimation in multivariate linear models. *Biometrika*, **99**, 687–702.

[61] Su, Z. and Cook, R.D. (2013), Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica*, **23**, 213–203.

[62] ter Braak, C.J.F. and de Jong, S. (1998), The objective function of partial least squares regression, *Journal of Chemometrics*, **12**, 41–54.

[63] Tso, M.K.S. (1981), Reduced-rank regression and canonical analysis. *JRSS-B*, **43**, 183–189.

[64] Tyler, D.E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics*, **9**, 725–736.

[65] Wen, X.M., Setodji, C.M. and Adekpedjou, A. (2009), A minimum discrepancy approach to multivariate dimension reduction via k-means inverse regression, *Statistics and its interface*, **2**, 503–511.

[66] Wold, H. (1966), Estimation of Principal Components and Related Models by Iterative Least Squares. *New York: Academic Press.*

[67] Yang, Y. (2005), Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika.* **92**, 934–950.

[68] Ye, Z. and Weiss, R.E. (2003), Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association*. **98**, 968–979.

[69] Yohai, V.J., Stahel, W.A. and Zamar R.H. (1991), A procedure for robust estimation and inference in linear regression, *Directions in Robust Statistics and Diagnostics, Part II, W. A. Stahel and S. W. Weisberg, Eds., Springer*, 365–374.

[70] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007), Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc. B* **69**, 329–346.

[71] Zhu, L.P, Zhu, L.X. and Wen, S.Q. (2010), On dimension reduction in regression with multivariate responses, *Statistica Sinica*, **20**, 1291–1307.

[72] Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, **15**, 265–286.